FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2014/2015

Trabajo de Fin de Máster

LA EDUCACIÓN, EL EMPLEO Y LOS HÁBITOS DE EMANCIPACIÓN DE LOS JÓVENES ESPAÑOLES

Censo de población y viviendas 2011

Alumna: Loubna Khalifi Chairi El Kammel

Tutor: Dr. Conrado Miguel Manuel García

Noviembre de 2015

UNIVERSIDAD COMPLUTENSE
MADRID

1 Introducción

Uno de los rasgos más distintivos de la juventud española, o al menos el más destacado en los últimos años, es que se trata de la generación mejor preparada, pero este aspecto tristemente ha ido acompañado a lo largo de los últimos años de crisis económica de las tasas más altas de desempleo juvenil. Dicho lo anterior, se sigue cumpliendo que las tasas de desempleo por lo general son más altas para la población con menor nivel educativo; según la Organización para la Cooperación y el Desarrollo Económicos (OCDE) alrededor del 74% de la población con educación superior tiene un empleo remunerado (El fenómeno del Becario, requeriría un estudio específico), en comparación con cerca del 47% para la población que no cuenta con educación postsecundaria. A pesar de ello, esta diferencia de 27 puntos porcentuales se encuentra muy por debajo de la media de la OCDE de 34 puntos porcentuales.

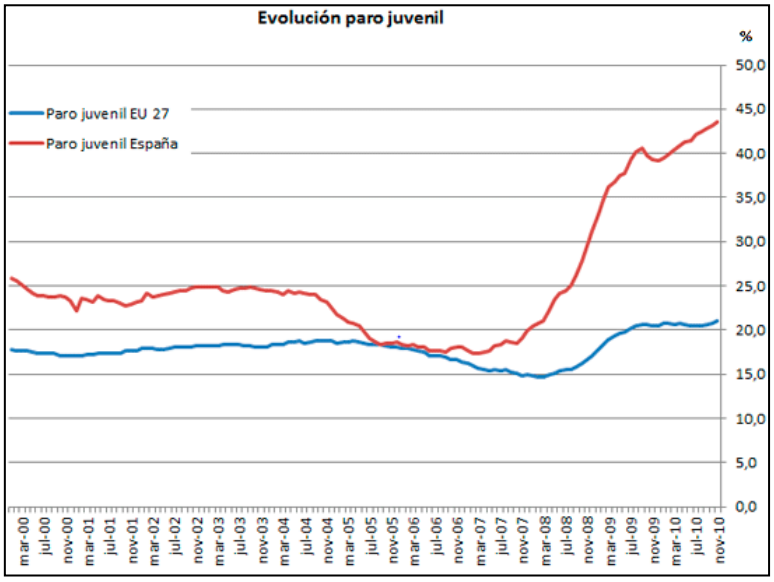


Figura 1.1 Evolución paro juvenil (Fuente Eurostat)

Y a pesar de que es cierto que las cifras generales del empleo, indican que sí hay mejores números en cuanto a la ocupación laboral en este último año, pero no así en su calidad profesional, especialmente en el caso de los jóvenes. Pues tal y como apuntan los resultados del último observatorio de la emancipación del Consejo de la Juventud de España (CJE) (correspondiente al cuarto trimestre de 2014) [1], cuyo objetivo es ofrecer un seguimiento de algunos elementos relacionados con el empleo y la vivienda que definen las condiciones de vida de la población joven en España, los tipos de contratos que más se han firmado son los de prácticas y formación. Tan solo en un año han incrementado un 58,01% su tasa interanual para los menores de 30 años.

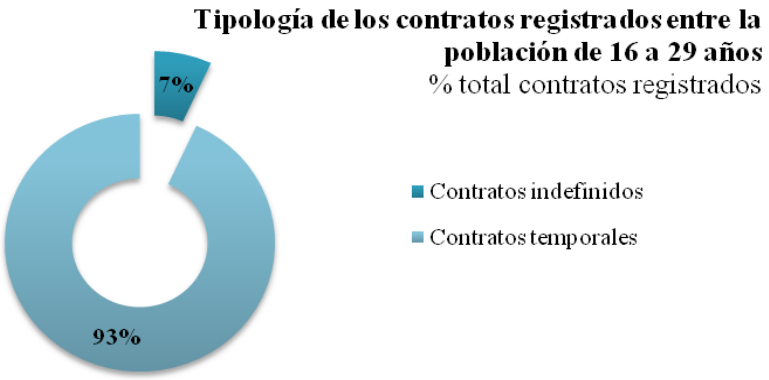


Figura 1.2 Gráfico de la tipología de contratos (Fuente CJE)

Este escenario es el que se están encontrando en los últimos años los jóvenes españoles cuando deciden plantear dar el gran salto de la emancipación. Por ello en los distintos estudios tanto a nivel europeo como a nivel de los países de la OCDE, España se encuentra entre los países con la media de edad de emancipación más alta, 29 años, 6 años más tarde que los franceses y tres años por encima de la media europea cuya media se sitúa en 26 años (25 para las mujeres, y 27 años para los hombres).

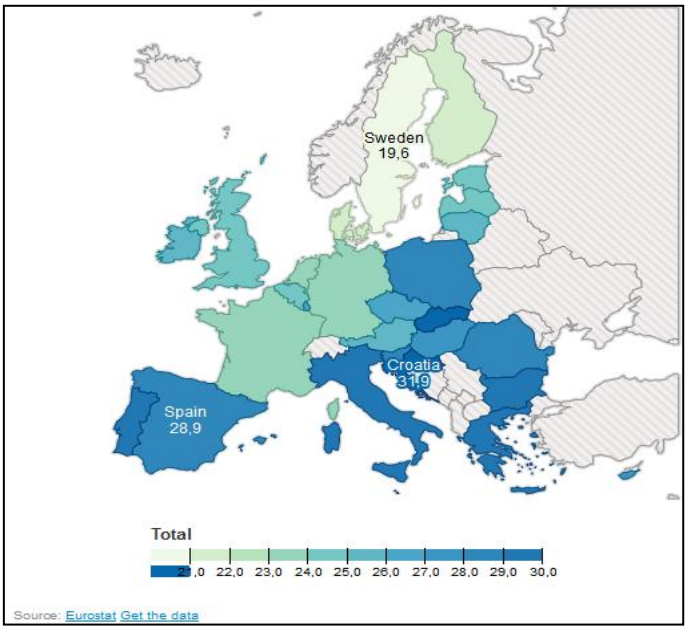


Figura 1.3 Edad de emancipación media (Fuente Eurostat)

En estos estudios se destaca que la creciente precariedad laboral es uno de los factores más importantes por los que se retrasa la edad de emancipación, pero casi al mismo nivel se suele situar también el mercado de la vivienda, que a pesar de haber sufrido un importante descenso en los últimos años de crisis, sigue arrastrando la inercia de los muchos años de burbuja inmobiliaria en los que se alcanzaron niveles de precios que ponían al alcance de muy pocos el poder acceder a la compra o alquiler de una vivienda con ciertas garantías de solvencia económica.

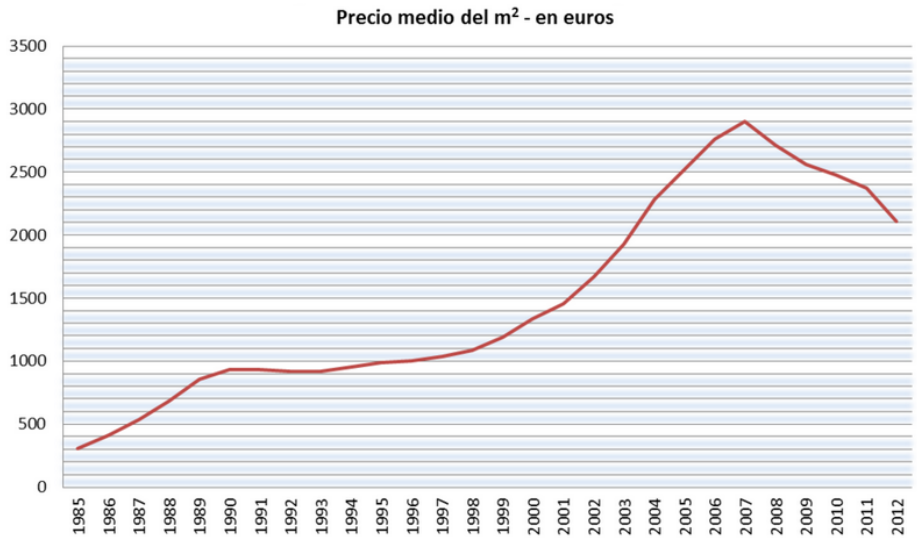


Figura 1.4 Evolución del precio de la vivienda en España
(Fuente Ministerio de Fomento)

2 Objetivos

Dado el escenario descrito, en el presente estudio se pretende realizar un acercamiento a las características socioeconómicas de los jóvenes españoles, incidiendo en las fases que definen sus condiciones de vida, como son el periodo de formación, la inserción en el mercado laboral, y acceso a la vivienda.

Para llegar a perfilar estos aspectos que caracterizan a los jóvenes españoles, desde un punto de vista cuantitativo y cualitativo, el trabajo se sustenta en la base de datos del Censo de Población y Vivienda de 2011 elaborado por el Instituto Nacional de Estadística (INE) [2], que a pesar de tratarse de resultados desarrollados hace cuatro años, son de total vigencia, ya que se trata del ejercicio de mayor despliegue de los que realiza el INE con una periodicidad de diez años, de hecho es el primero en desarrollarse según los estándares europeos, consiguiendo resultados homogéneos, a nivel de escalas, que permiten realizar comparativas con los países miembros de la unión. La muestra recogida representa aproximadamente el 12% (4.075.295) de la población total.

Dentro del ámbito del análisis estadístico, el volumen de datos al que nos enfrentamos, nos lleva directamente a recurrir a las técnicas de minería de datos para abordar los objetivos planteados. De hecho, la generación de un buen modelo de minería de datos se convierte en sí en un objetivo, que forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de los distintos modelos para darles respuesta.

Centrando el foco en los diferentes estudios que se han realizado a lo largo del trabajo, se podrían resumir en los siguientes puntos:

- Describir la muestra de censo de población y vivienda 2011, y nuestra población de interés, los jóvenes entre 22 y 30 años.
- Analizar las comunidades autónomas en base a los factores macroeconómicos, para su posterior agregación.
- Conocer el nivel educativo que presentan los jóvenes españoles, en función de sus características socio-demográficas y de la formación y ocupación de sus padres.
- Estudiar los atributos que caracterizan a los jóvenes emancipados.
- Relacionar la ocupación de los jóvenes universitarios con los estudios realizados.
- Analizar el régimen de tenencia de la vivienda entre los jóvenes emancipados.

Este documento se ha estructurado en los siguientes bloques de contenidos, que reflejan las sucesivas fases del estudio:

1. En la primera y la segunda sección se presentan la introducción y los objetivos marcados.
2. En la tercera sección se expone la metodología estadística empleada.
3. En la cuarta sección se describe la fuente de datos.
4. En la quinta sección se aplica la metodología SEMMA con todas sus fases.
5. En la sexta sección se analiza la influencia del nivel educativo de los progenitores en los estudios de los hijos por medio de un modelo logístico ordinal.
6. En la séptima sección se aborda el tema de la ocupación de los jóvenes universitarios.
7. En la octava sección se estudia mediante un modelo logístico la probabilidad de emancipación.
8. En la novena sección se desarrollarán las técnicas de aprendizaje automático para pronosticar el régimen de tenencia entre los jóvenes emancipados.
9. En la décima sección se presentan las conclusiones del estudio.
10. En la undécima sección se indica la bibliografía utilizada.

Por último, señalar que la herramienta informática utilizada para el tratamiento estadístico de la base de datos han sido los paquetes SAS Enterprise Miner 13.1, SAS Base 9.4 e IBM SPSS v19, versiones para Windows.

3 Metodología Estadística

Como se ha comentado en el apartado anterior, se hace necesario recurrir a técnicas estadísticas que nos ayuden a comprender mejor los datos, de tal forma que las decisiones que se tomen estén fundamentadas. En el presente bloque se resume la metodología estadística seguida en la elaboración de este trabajo.

3.1 Análisis Cluster

Mediante esta técnica se persigue identificar o agrupar individuos u objetos que son similares con respecto a un criterio, en tipos o grupos que internamente son homogéneos y que entre sí son heterogéneos. Esta técnica se empleará para agrupar a las comunidades autónomas con respecto a los datos macroeconómicos.

3.2 Modelos de Elección Discreta

Los modelos de elección discreta son modelos estadísticos en los que se desea conocer la relación entre una variable dependiente discreta, y una o más variables explicativas independientes, ya sean cualitativas o cuantitativas. La tipología más empleada de estos modelos son los denominados Logit o modelos de regresión logística, siendo la ecuación inicial del modelo de tipo exponencial, si bien su transformación logarítmica (Logit) permite su uso como una función lineal, en los casos mencionados.

A lo largo del estudio, se aplicara por un lado regresión logística ordinal para el análisis de nivel educativo de los jóvenes. Y por otro lado, para averiguar los factores que influyen en la independencia de los jóvenes, así como en el régimen de tenencia de los jóvenes emancipados se empleará la regresión logística binaria.

Se presenta a continuación la ecuación para modelo logístico ordinal, las comparaciones de este modelo se llevaran a cabo en cada punto de corte de la escala ordinal.

$$y^* = x'\beta + \varepsilon$$

$$y = \begin{cases} 1 \quad si \quad y^* \leq \mu_1 \\ 2 \quad si \quad \mu_1 \leq y^* \leq \mu_2 \\ \vdots \\ j \quad si \quad \mu_{j-1} \leq y^* \end{cases}$$

- y^* *la variable dependiente no observada*
- y_j *la variable depentiente observada, siendo j los niveles de dicha variable*
- μ_j *umbrales que hay que estimar*

- $x\ vector\ de\ variables\ explicativas$
- $\beta\ parametros\ a\ estimar$

3.3 Análisis Factorial de Correspondencias Simples

El análisis factorial de correspondencias simples es una técnica estadística multivariante de interdependencia, entre dos variables cualitativas. Esta técnica busca establecer en lugar de relaciones causales entre las dos asociaciones entre sus categorías. Se llevará a cabo la técnica para examinar las relaciones que puedan existir entre los estudios cursados y el cargo que ocupan los jóvenes, siempre y cuando haya asociación entre las categorías de las dos variables objeto de estudios.

3.4 Redes Neuronales

Una de las grandes diferencias de la red neuronal frente a otro tipo de aplicaciones consiste en que no son algorítmicas, no se emplea una fórmula matemática explícita. A la hora de programarlas no se indica un patrón fijo de instrucciones a realizar, sino que la propia red neuronal es capaz de elaborar sus propias "reglas" a fin de hallar la mejor respuesta a una entrada determinada.

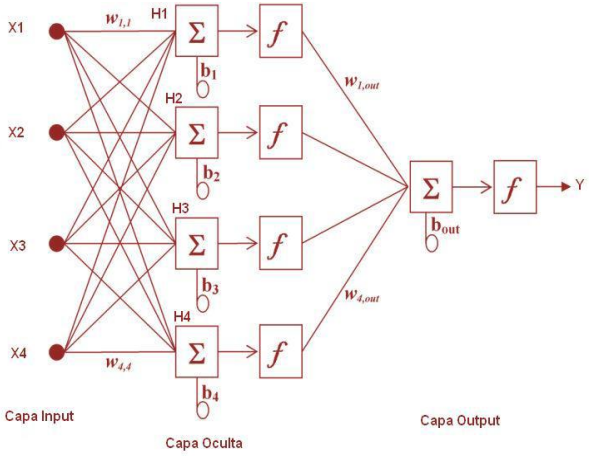


Figura 3.1 Funcionamiento red neuronal

En la Figura 3.1 se presenta un perceptrón multicapa, en el cual se consideran: las capas de datos Input (entradas, estímulos), capa de nodos ocultos (puede ser más de una) y finalmente la capa de salida. La ecuación de modelo de red neuronal sería:

$$y = f\left(w_{j,out}H_j + \ \ldots + w_{1,out}H_1\right) + \ b_{out} = f\left(\sum_{j}^{m} w_{j,out}H_j \ + \ b_{out}\right) = f(a)$$

$x_i\ Conjunto\ de\ entradas\ i = 1 \ldots n$

$w_{i,j}\ Pesos\ capa\ input\ con\ la\ capa\ oculta\ \ i = 1 \ldots n \quad j = 1 \ldots m$

$w_{j,out}$ $Pesos\ capa\ output\ \ j = 1 \ldots m$

$H_j = w_{1j}x_1 + w_{2j}x_2 + \cdots + w_{nj}x_n + b_j = \sum_{i=1}^{n} w_{ij}x_i$. Es una combinación lineal de los valores de entrada y los pesos.

3.5 Árboles de Clasificación

El árbol de decisión es una de las herramientas más útiles y utilizadas para la toma de decisiones adecuadas teniendo varias alternativas posibles de acción. El nombre de árbol de decisión proviene de la forma que adopta el modelo, semejante a un árbol. Está formado por múltiples nodos cuadrados, que representan los puntos de decisión, y de los cuales surgen ramas que representan las distintas alternativas. Mediante los árboles se aplicaran los algoritmos Bagging y Random Forest que se detallarán en el apartado 9.4.

Las redes neuronales y los arboles de clasificación se emplearán para pronosticar el régimen de tenencia de los jóvenes emancipados, ya que el objetivo en este caso no es el de explicar o cuantificar las probabilidades asociadas a cada factor, sino el de utilizar las técnicas de aprendizaje automático y averiguar que método clasifica mejor los datos.

4 Población Objetivo

4.1 Censos de Población y Viviendas 2011

Los Censos de Población y Viviendas son la operación estadística de mayor envergadura que el INE realiza cada diez años. Con estos censos se persigue la recogida de información sobre muchos aspectos que nos ayudan a conocer mejor nuestra sociedad, como pueden ser:

- La estructura de la población
- Las formas de convivencia
- La movilidad geográfica
- La relación con la actividad y la ocupación
- Las características de las viviendas

Esta información es accesible y es utilizada tanto por organismos públicos, empresas privadas, investigadores y ciudadanos.

La información utilizada en este estudio es la correspondiente al censo de 2011. A continuación se presenta un resumen sobre la variación de los parámetros básicos de este censo respecto al último realizado en 2001:

	Censo 2011	Censo 2001	Variación (%)
Población total	46.815.916	40.847.371	14,6
Hombres	23.104.303	20.012.882	15,4
Mujeres	23.711.613	20.834.489	13,8
Población en colectivos	444.101	233.347	90,3
Edificios	9.814.785	8.661.183	13,3
Viviendas (total)	25.208.623	20.946.554	20,3
Viviendas vacías	3.443.365	3.106.422	10,8
Hogares	18.083.692	14.187.169	27,5

Tabla 4.1 Resumen comparativo

También se representa la distribución de la muestra según los principales ejes de segmentación que serán empleados en el estudio (Edad, Sexo, Nacionalidad):

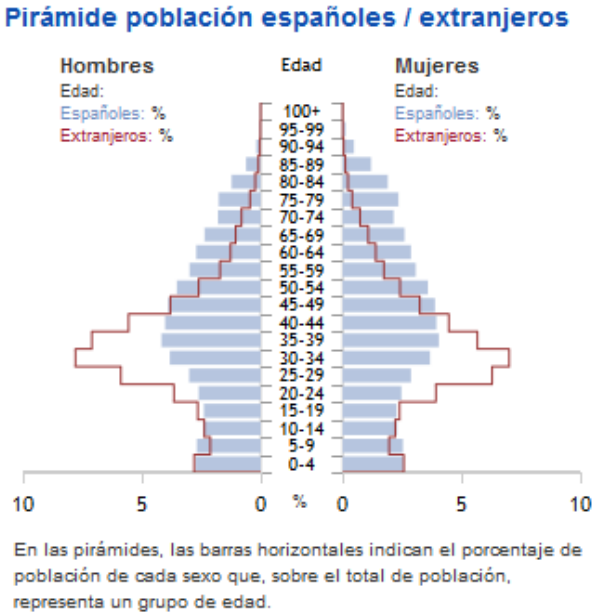


Figura 4.1 Pirámide población censo 2011

Este estudio se centra en la submuestra correspondiente a la población de jóvenes entre 22 y 30 años cuya selección se explicará más en detalle en el apartado de Fase de Muestreo 5.1 de la **metodología SEMMA**.

4.2 Escenario Macroeconómico

En general modelizar y explicar el comportamiento de las variables objeto de estudio constituye una tarea difícil de abordar, sobre todo cuando se tiene poca información relativa a estas, o cuando a pesar de la riqueza de la información disponible, esta no engloba todos los factores que influyen en la variabilidad de las respuestas.

Ante esta situación, en este trabajo se ha optado por completar la información que proporciona el censo de población y vivienda con variables macroeconómicas desagregadas a nivel comunidad autónoma y referida al periodo de referencia del censo.

Con ello se podrá construir modelos mixtos, combinando datos socio-demográficos y microeconómicos, con el uso de las variables macroeconómicas (Tabla A1.1, Anexo 1). A continuación se presenta el gráfico de los datos macroeconómicos por comunidad autónoma. Los valores corresponden a la variación interanual 2011/2010, a excepción de la tasa de paro.

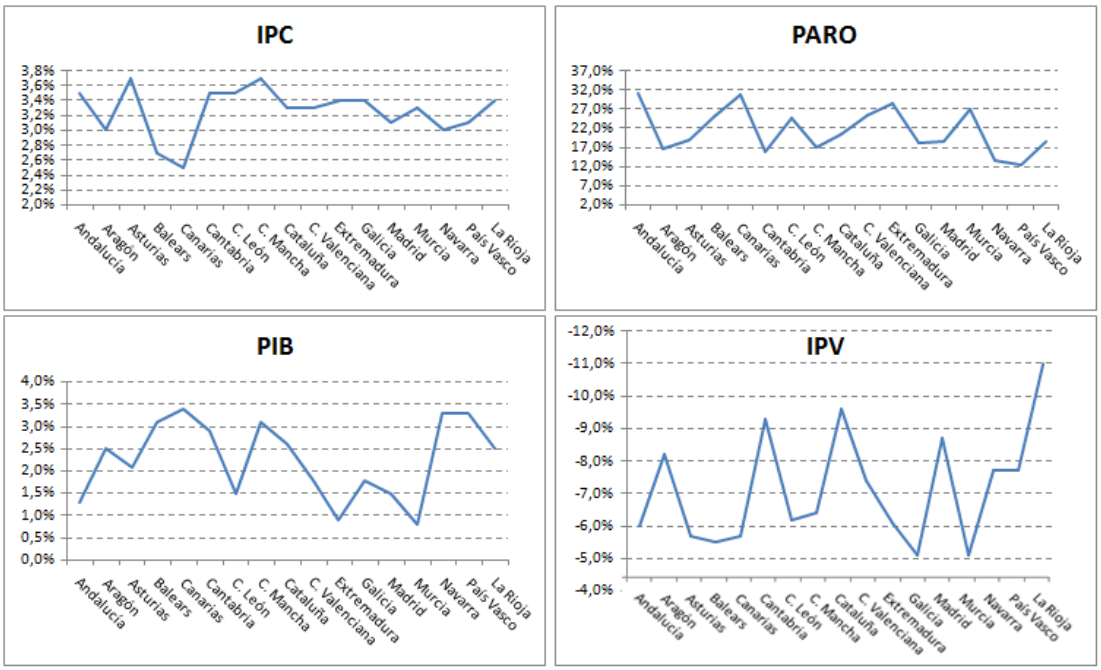


Figura 4.2 Datos macroeconómicos por Comunidad autónoma

- La tasa de paro empleada en el estudio, se obtiene mediante la encuesta de población activa, mediante la cual se alcanzan los resultados de la fuerza de trabajo y de sus diversas categorías (ocupados, parados), así como de la población ajena al mercado laboral (inactivos).
- El Producto Interior Bruto (PIB) es un indicador económico que refleja la producción total de bienes y servicios asociada a un país durante un determinado periodo de tiempo.
- El Índice de Precios de Vivienda (IPV) tiene como objetivo medir la evolución de los precios de compraventa de las viviendas de precio libre, tanto nuevas como de segunda mano, a lo largo del tiempo.
- El índice de precios de consumo (IPC) es una medida estadística de la evolución de los precios de los bienes y servicios que consume la población.

Cabe mencionar que la información relativa a estos datos macroeconómicos no se tendrá en cuenta a la vez que la comunidad autónoma en los modelos que ajustaremos al existir multicolinealidad, sino que se irá probando las distintas combinaciones hasta alcanzar el mejor modelo predictivo.

5 Metodología SEMMA

Una vez presentados los datos que se van a tener en cuenta en las distintas fases del estudio, se hace necesario definir qué estrategia se va seguir para abordar y tratar los datos de los que se dispone, o lo que es lo mismo cómo se va estructurar el proceso de minería de datos.

Para llevar a cabo este cometido existen numerosas metodologías, pero las que presentan mayor aceptación entre los desarrolladores de modelos de minería de datos son CRISP (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, Asses), en este trabajo se ha optado por esta última alternativa, ya que como se ha comentado anteriormente SAS será el Software Estadístico utilizado en el análisis de los datos, y la metodología SEMMA está adaptada a los procesos ya implementados en esta herramienta ya que fue desarrollada en 1998 por el **SAS Institute**.

Una de las características de esta metodología es que la implementación de las fases en las que se estructura (selección, exploración, modificación y modelado) no es rígida, es decir, no es necesario terminar una de sus fases antes de comenzar otra.

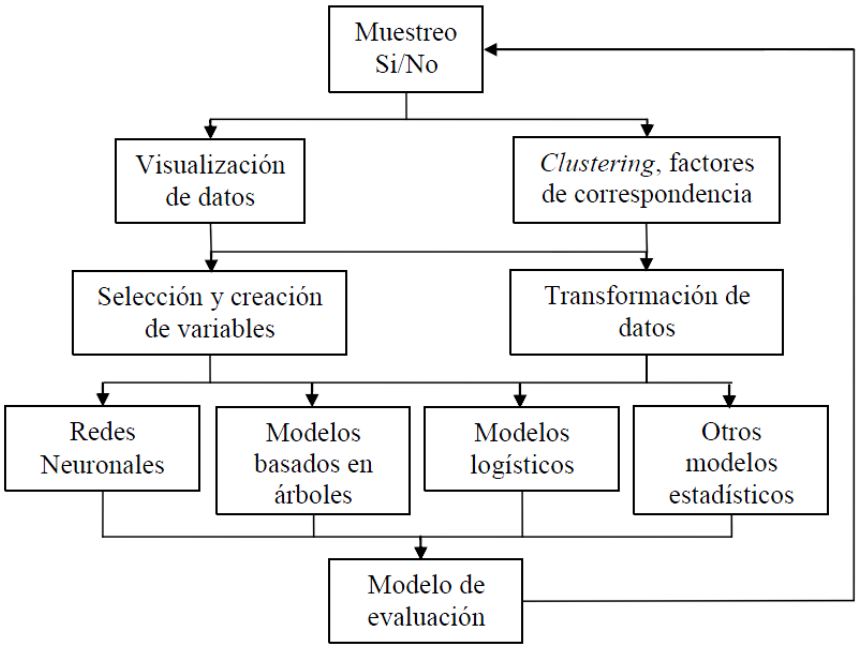


Figura 5.1 Metodología SEMMA

5.1 Fase Muestreo

En esta primera fase de la metodología, se define el ámbito de la muestra de trabajo y se garantiza su representatividad, este último punto es crucial ya que los resultados que se extraigan de la muestra tienen que ser aplicables a toda la población objetivo del estudio.

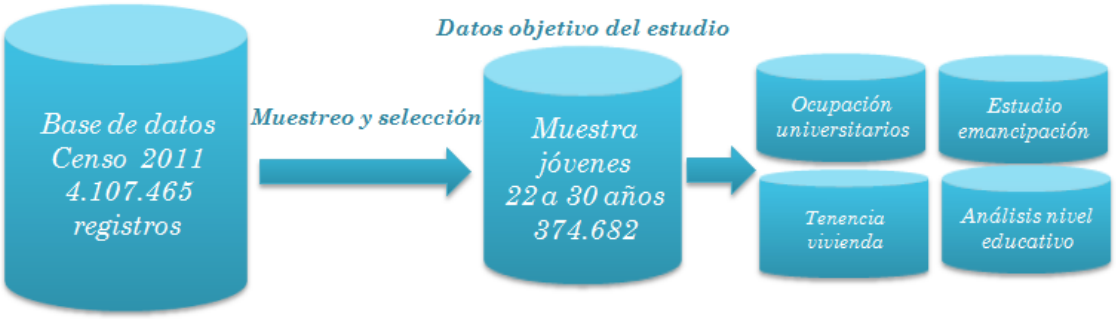


Figura 5.2 Fase de muestreo

Por otro lado, con la fase de muestreo se consigue a su vez la reducción de los tiempos de ejecución de algunas técnicas, como pueden ser las Redes Neuronales cuyo tiempo computacional es muy lento, sin afectar a la robustez de los resultados obtenidos.
Volviendo a nuestra fuente de información, el censo de población y vivienda 2011 realizado por el INE, que combina el uso de registro administrativo (Padrón municipal) con la información de una gran muestra formada por 1.621.643 hogares y 4.107.465 personas.

En este trabajo se parte de esta última muestra extrayendo la submuesta formada por los jóvenes de 22 a 30 años, que contienen a 374.682 personas, un 9.12%. La selección de esta franja de edad ha sido condicionada en primer lugar por los estudios que se quieren emprender, abordando los estudios que han cursado los jóvenes, su inserción laboral y su emancipación, y en segundo lugar por tratarse del intervalo de edad utilizado en multitud de trabajos que abordan estas temáticas.

Las comunidades autónomas de Ceuta y Melilla quedan excluidas en el presente trabajo al presentar un porcentaje irrelevante una vez seleccionada la muestra de jóvenes, exactamente un 0.14% y 0.16% respectivamente.

5.2 Fase Exploración

Una vez definido el ámbito de la muestra, se procede a explorarla para conocer en detalle de que información se dispone y en qué estado se encuentra, y con ello adelantar los tratamientos que se deberían realizar en la fase de modificación.

En esta exploración se realizan análisis descriptivos de las variables, tanto univariantes como bivariantes cruzándolas con las distintas variables objetivo de estudio, así como procesos de detección de valores desinformados o atípicos, aunque respecto de este último punto hay que decir que al tratarse de una muestra del censo nacional, está ya ha

sufrido varias fases de depuración y por lo tanto prácticamente no presenta fallos en la información como suele ocurrir habitualmente.

Figura 5.3 Fase de exploración

La base de datos de la que disponemos contiene información de 139 variables repartidas en distintos bloques:

• Identificación	• Datos del hogar
• Datos individuales	• Datos del padre
• Datos de la vivienda	• Datos de la madre
• Datos del edificio	• Datos de cónyuge o pareja
• Datos de parentesco	• Datos de núcleo

Tabla 5.1 Bloques cuestionario censo 2011

Para los distintos apartados iremos seleccionado los bloques que estén relacionados con el tema a estudiar, para así facilitar el manejo de la tabla así como el tiempo de ejecución. A continuación pasaremos a la fase de explorar la muestra.

En la tabla 5.2 se muestra la distribución de sexo, donde se aprecia que se tienen aproximadamente la misma proporción de mujeres que de varones.

Sexo	Frecuencia	Porcentaje
Hombre	188899	50.42
Mujer	185783	49.58

Tabla 5.2 Distribución sexo

En cuanto a la edad (Tabla 5.3), el intervalo de estudio se centra desde los 22 hasta los 30 años, se tiene que la edad media se sitúa en 26.1 años con una desviación típica de 2.61 años sobre la media, estando la mediana en 26 años.

Variable de análisis : Edad						
N	**Mínimo**	**Media**	**Devstd**	**Mediana**	**Moda**	**Máximo**
374682	22	26.1	2.61	26	30	30

Tabla 5.3 Descriptivos variables edad

En lo que respecta al nivel académico (Figura 5.4), los niveles educativos que acumulan mayores porcentajes son los estudios universitarios, seguidos de las enseñanzas secundarias y bachillerato, por lo que podría decirse que los jóvenes poseen unos niveles elevados de formación.

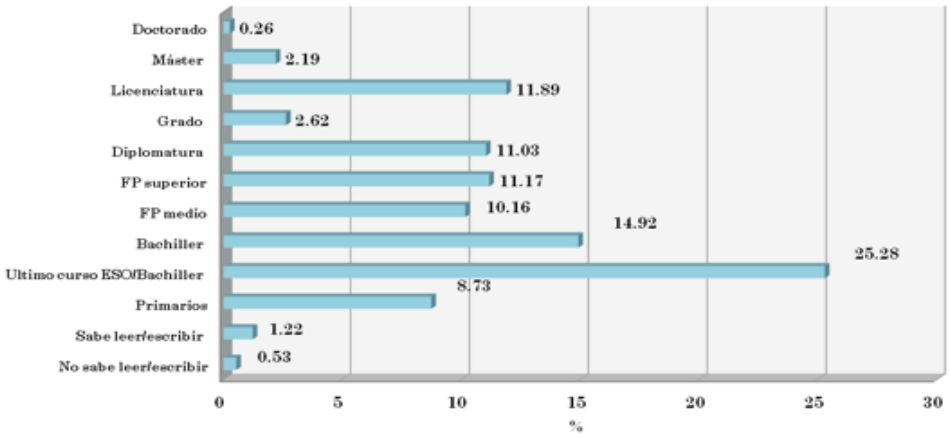


Figura 5.4 Nivel de estudios completados

En la Figura 5.5 vemos la relación de los jóvenes con el ámbito laboral, en primer lugar se observa que un 52.25% de los jóvenes tiene empleo, seguido de un 24.09% de los jóvenes que se encuentran en situación de desempleo, pero que ha estado empleada antes, frente a un 6.81% que están inactivos por primera vez.

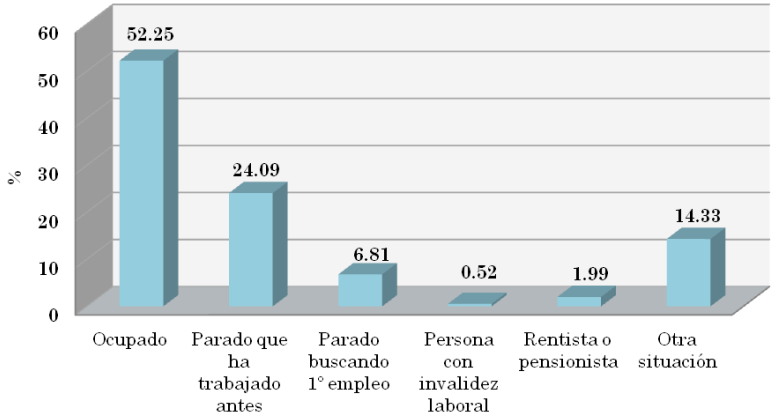


Figura 5.5 Relación con la actividad

En la siguiente figura, se aprecia que hay más mujeres con estudios universitarios entre la población de jóvenes de 22 a 30 años de edad, estas diferencias también se manifiestan en los estudios primarios y secundarios obligatorios siendo la proporción predominante la de los hombres. En cuanto a la formación profesional superior y bachiller no se aprecian diferencias significativas.

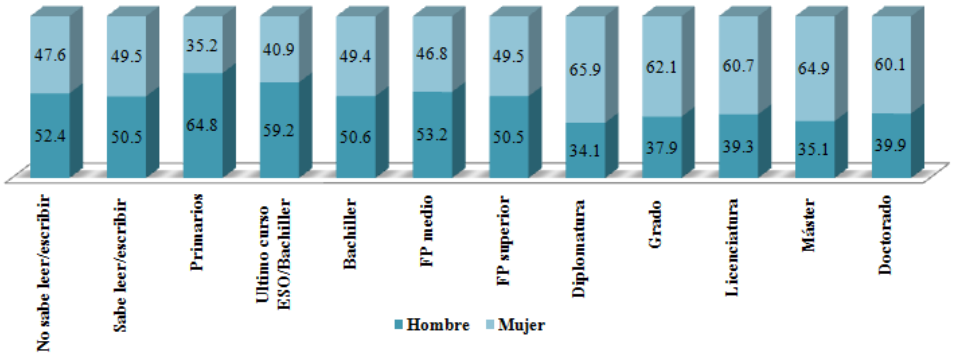


Figura 5.6 Nivel educativo por sexo

5.3 Fase Modificación

Tomando como base los resultados obtenidos en la fase de exploración anterior, en esta etapa nos centramos en la selección y transformación de las variables que serán el input para la construcción de nuestros modelos. Entre otras tareas a realizar destacan: la imputación de valores desinformados e atípicos, la reducción de dimensión, la creación de nuevas categorías que aglutinen aquellas sin representatividad.

Figura 5.7 Fase modificación

Se ha estructurado este apartado en dos bloques, en primer lugar se expondrán las recategorizaciones y la creación de nuevas variables. Y en segundo lugar, se ha planteado agrupar las comunidades autónomas en base a las variables macroeconómicas presentadas en el apartado 4.2.

5.3.1 Recategorizaciones

Una vez que se hayan explorado las variables objeto de estudio en el trabajo, se ha visto que hay algunas categorías irrelevantes en las que apenas representan el 5% de los datos, y por otro lado hay variables que presentan un número extenso de niveles.

Variable antigua	**Varibale nueva**
Código país de nacionalidad (Código de los 198 Países + Apátridas)	***Nacionalidad*** (Española / Extranjera)
Código de provincia (52 provincias)	***Comunidad Autónoma*** (17 Comunidades)
Tamaño municipio : tamaño de municipio <=2.000, Si 2.001 <=tamaño de municipio <=5.000, Si 5.001 <=tamaño de municipio <=10.000, Si 10.001 <=tamaño de municipio <=20.000, Si tamaño de municipio >20.000	***Área residencial*** (Municipios <=20.00 / Municipios > 20.000)
Código de ocupación: valores CNO a 2 dígitos (90 categorías)	***Código de ocupación al nivel superior*** un dígito (10 categorías)
Número de habitaciones : valores de 1 a 30 habitaciones	***N habitaciones:*** De 1 a 6, más de 6

Tabla 5.4 Creación de nuevas variables

A tenor de lo que se ha observado en la fase de exploración, se ha procedido a crear las nuevas variables que se muestran en la Tabla 5.4, y por otro lado las variables nivel educativo y la actividad que mostraban categorías con poca representatividad, se han agrupado de la de la siguiente forma. En el anexo se presentan el resto de las agrupaciones (Tabla A2.1 hasta Tabla A2.3, Anexo 2).

Estudios	Frecuencia	Porcentaje
Primarios	39283	10.48
Ultimo curso ESO/Bachiller	94704	25.28
Bachiller	55888	14.92
Fp	79946	21.34
Universitarios	104861	27.99

Tabla 5.5 Nueva variable estudios

Actividad	Frecuencia	Porcentaje
Ocupado	195788	52.25
Parados	115805	30.91
Otra situación	63089	16.84

Tabla 5.6 Nueva variable actividad

Estas agrupaciones se analizarán con cada modelo que vayamos ajustar, para cerciorarnos de que la agrupación este bien hecha.

5.3.2 Agrupaciones de las comunidades autónomas

En este apartado se procede a tratar la información auxiliar que se añadirá a la muestra del censo, formada por las variables macroeconómicas PIB, Paro, IPC e IPV. Se dispone para cada una de las 17 comunidades autónomas con representación en la muestra de la variación interanual 2011/2010 de estas variables, y en base a estas mismas variables se ha observado que algunas comunidades presentan valores muy próximos, por lo que se ha visto interesante estudiar si se podría realizar una agrupación.

Para realizar esta agrupación se recurre al análisis cluster, con el que se intentará formar *conglomerados* que cumplan que, los objetos dentro de cada conglomerado, son similares entre sí (alta homogeneidad interna) y diferentes a los objetos de los otros conglomerados o clusters (alta heterogeneidad externa).

Las medidas de distancia del análisis Cluster son sensibles a la diferencia de escalas o de magnitudes hechas entre variables, en consecuencia es necesaria la estandarización de datos para evitar que las variables con una gran dispersión tengan un mayor efecto en la homogeneidad. En nuestro caso nos vemos obligados a estandarizar ya que las variables no están medidas en la misma unidad.

Variable	N	Media	Devstd	Mínimo	Máximo
IPC	17	3.26	0.33	2.50	3.70
PARO	17	19.04	5.43	10.90	28.96
PIB	17	2.26	0.87	0.80	3.40
IPV	17	7.14	1.75	5.10	11.00

Tabla 5.7 Estadísticos descriptivos de las variables Macroeconómicas

Se realiza el análisis cluster aplicando el algoritmo de clasificación ‘enlace por mínima varianza’ que unirá en un nuevo nivel la pareja de grupos que produzca el mínimo incremento en la varianza residual.

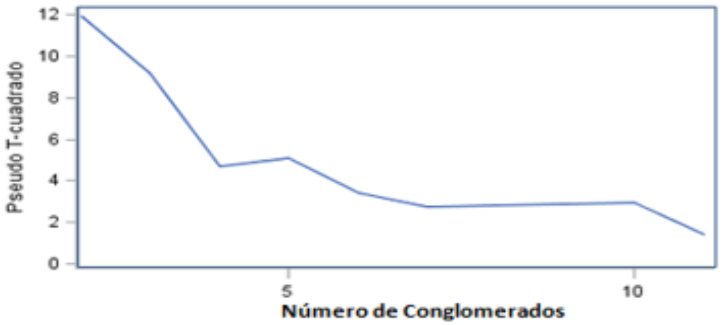


Figura 5.8 Criterio para número de clúster del estadístico Pseudo T-Squared.

Observando la tabla 5.8 junto con la figura 5.8 se aprecia un máximo local en 5 clusters, lo que aconsejaría tomar el anterior agrupamiento es decir 6 clusters, este agrupamiento tiene por R^2= 0.874, lo que se traduce en un 87.4% de proporción de variabilidad explicada por los clusters.

Historia de conglomerado							
Nº clusters	Conglomerados unidos		Frec	R-cuadrado semiparcial	R-cuadrado	Estadístico pseudo F	T-cuadrado pseudo
16	Navarra	País Vasco	2	0.001	0.999	75.700	.
15	Extremadura	Murcia	2	0.004	0.995	31.000	.
14	Cantabria	Cataluña	2	0.007	0.989	20.400	.
13	C Mancha	C. Valenciana	2	0.008	0.981	16.900	.
12	Asturias	Galicia	2	0.009	0.972	15.700	.
11	CL14	La Rioja	3	0.009	0.963	15.400	1.400
10	Andalucía	CL15	3	0.011	0.952	15.300	3.000
9	Aragón	Madrid	2	0.012	0.940	15.600	.
8	Baleares	Canarias	2	0.016	0.924	15.600	.
7	CL12	C León	3	0.024	0.900	15.000	2.800
6	**CL10**	**CL13**	**5**	**0.0259**	**0.874**	**15.3**	**3.4**
5	CL9	CL11	5	0.047	0.827	14.400	5.100
4	CL5	CL16	7	0.071	0.756	13.400	4.700
3	CL6	CL7	8	0.125	0.631	12.000	9.200
2	CL4	CL8	9	0.277	0.354	8.200	12.000
1	CL3	CL2	17	0.354	0.000	.	8.200

Tabla 5.8 Historial de los clusters

El análisis de conglomerados reveló la existencia de seis grupos como se puede apreciar en el dendograma (Figura A2.1, Anexo 2), este último es una valiosa herramienta visual que puede ayudar a decidir el número de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en que se van anidando los cluster. Aunque la decisión tomada se basa en los resultados numéricos.

Una vez determinado el número de clusters, se va a caracterizar cada uno de ellos. En la tabla 5.9, se presentan los valores medios de cada una de las variables descriptivas dentro de cada cluster, y se enfrenta con la media nacional con el fin de destacar los atributos más relevantes de cada uno de los clusters.

Variables Macroeconómicos	Nacional	Clusters																	
		Navarra	P. Vasco	Andalucía	Extremadura	Murcia	C. Valenciana	C. Mancha	Cantabria	Cataluña	La Rioja	C. León	Asturias	Galicia	Aragón	C. Madrid	Baleares	Canarias
Variación 11/10 IPC	3.30	3.05		3.40					3.40			3.60			3.05		2.60	
Tasa de desempleo	22.80	11.27		24.29					16.18			16.05			15.90		25.59	
Variación 11/10 PIB	2.10	3.30		1.26					2.67			2.33			2.00		3.25	
Variación 11/10 precio vivienda	7.40	7.70		6.16					9.97			5.73			8.45		5.60	

Tabla 5.9 Características de los clusters

Atendiendo a los datos expuestos en la tabla, podemos describir de manera breve las características de cada cluster, cabe mencionar que se ha realizado el análisis multivariado de la varianza para contrastar la hipótesis nula de igualdad de medias de los seis grupos de clusters en el conjunto de las cuatro variables macroeconómicas (MONOVA). Los resultados arrojan diferencias estadísticamente significativas entre las medias de los grupos. En el anexo se muestran los resultados del análisis, así como la constatación de los supuestos del estudio (Tabla A2.4 hasta Tabla A2.7, Anexo 2).

- **Cluster 1:** Es el formado por las comunidades País vasco y Navarra que presentan valores en torno a la media nacional en las variaciones del IPC y el precio de la vivienda, en cambio la tasa de desempleo se sitúa muy por debajo de la media nacional .
- **Cluster 2:** Las cinco comunidades que lo conforman registran altas tasas de paro, y obtienen peores registros en términos de variación interanual del PIB.
- **Cluster 3:** Destaca por incluir a las comunidades autónomas que mayor descenso registraron en el precio de la vivienda, aproximadamente una media del 10% respecto al 2010. En cuanto a los demás indicadores señalar que las variaciones del IPC y el PIB están por encima de la media nacional.
- **Cluster 4**: Este conglomerado abarca Galicia, Castilla y León, y Asturias. Muestra la misma estructura que el anterior cluster salvo por la variación del precio de la vivienda, ya que en este cluster solo se presenta una bajada del 5.73% en media.
- **Cluster 5:** Las comunidades de Madrid y Aragón que forman este grupo destacan por presentar valores entorno a la media nacional, salvo en la tasa de paro que se encuentra por debajo de la media nacional.
- **Cluster 6:** Y por último las Islas Canarias y Baleares que forman este grupo se diferencian de los cinco anteriores por tener la tasa de paro más alta y por presentar la menor variación media del IPC respecto a los datos nacionales.

Por último cabe destacar que esta agrupación no será utilizada en todos los modelos, ya que en los casos en los que se considere que el mayor detalle en la variable proporciona mejores resultados, se considerará sin la nueva agregación definida.

5.4 Fase de Modelar y Evaluar

Finalmente la fase que culmina este proceso: se inicia el proceso de selección y desarrollo de los modelos que se van a implementar. El proceso de selección, aunque a priori parezca directo, ya que viene limitado por las variables que se quieran modelizar, no es sencillo por el hecho de que actualmente se disponen de una amplia gama de técnicas y herramientas, cada una con sus pros y contras, y el hecho de acertar en esta elección condiciona todo el proceso posterior.

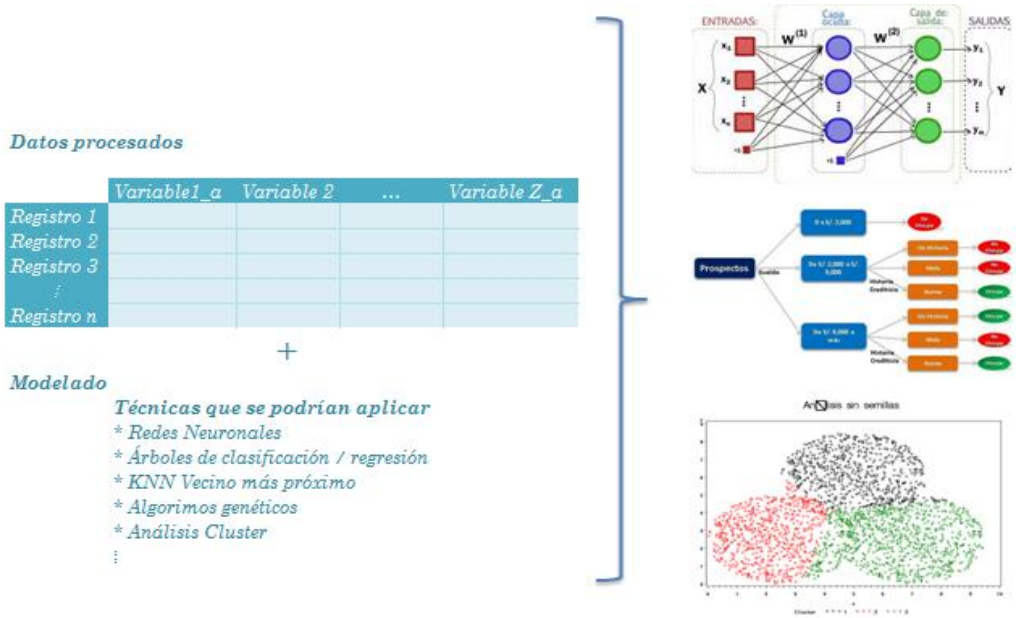


Figura 5.9 Fase modelado

Por último, una vez elegidos los modelos entre los distintos desarrollados, se evalúa su calidad y consistencia contrastando con otros modelos estadísticos o mediante validación cruzada (Apartado 9.1.2) antes de ser usado en la realidad.
La puesta en práctica de lo explicado en esta fase se detalla en cada uno de los apartados siguientes donde se desarrollan los distintos modelos predictivos y descriptivos.

Cabe mencionar que se han desarrollado códigos para el tratamiento de la muestra y para el ajuste de los modelos. También se ha hecho uso de las macros SAS desarrolladas por el Profesor *Javier Portela García-Miguel*, facilitadas en la asignatura de máster *'Redes Neuronales y Algoritmos Genéticos'* para la implementación de los modelos de aprendizaje automático.

6 Determinantes del nivel de estudios de los jóvenes

El abandono prematuro de la educación en España ronda niveles muy altos en comparación con el resto de países europeos, presentando así tasas (datos 2011) que oscilan entre 18% y 40% según la comunidad autónoma. Estos datos son superados únicamente por algunas regiones de Turquía como se puede apreciar en la Figura 6.1.

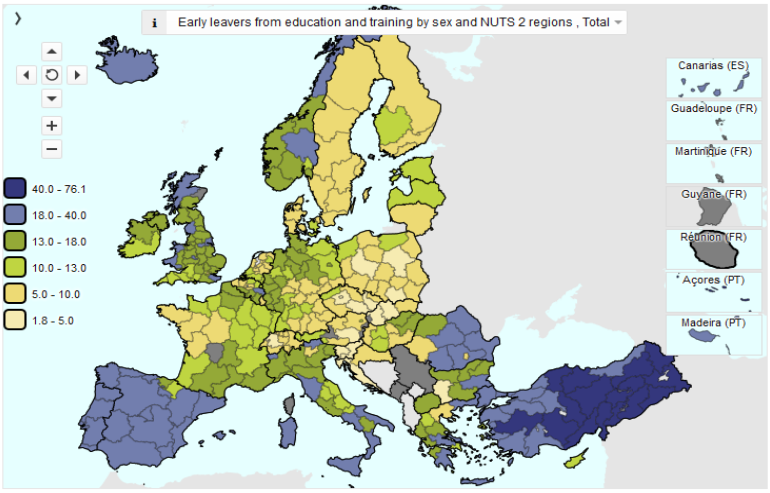


Figura 6.1 Tasa de abandonos prematuros de la educación y la formación (fuente Eurostat)

Dentro de este contexto, vamos a describir los resultados más detalladamente a nivel nacional en los últimos años antes de ahondar en la técnica estadística empleada.

En el siguiente gráfico se muestra el porcentaje de población que ha completado como máximo la primera etapa de la educación secundaria y no sigue ningún estudio o formación. Se aprecia como la serie histórica presenta una tendencia decreciente hasta el 2001, que observamos como empieza a aumentar ligeramente en los años de bonanza económica. Este comportamiento se mantiene aproximadamente constante hasta el 2009, a partir de entonces, se aprecia una clara tendencia a la baja, alcanzando los valores más bajos en la serie histórica. Este hecho podría ser consecuencia de las altas tasas de desempleo que sufre España desde entonces, lo que hace que los jóvenes no abandonen los estudios para incorporarse al mercado laboral, o incluso los reanuden.

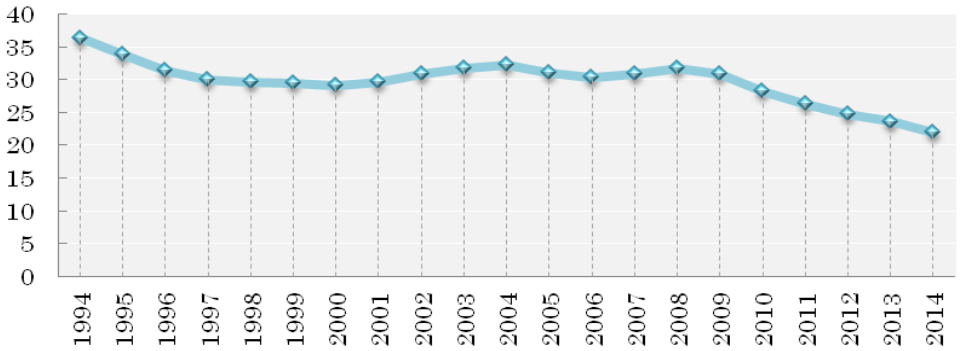


Figura 6.2 Tasa de abandono escolar prematuro de la población total de 18 a 24 años

6.1 Modelo de elección discreta. Regresión logística ordinal

Estudios realizados en diferentes países, destacan la influencia que tiene el nivel de estudios de los progenitores en la educación de los hijos, así como otras variables socio-demográficas del entorno de los jóvenes que iremos detallando a lo largo del apartado.

Para poner de manifiesto lo que los estudios desvelan, se va a ajustar un modelo de regresión logística ordinal (Apartado 3.2) con la variable nivel de estudios, con la finalidad de determinar qué factores influyen en tener un nivel de estudios u otros. Para llevar a cabo el análisis, se ha centrado el foco en los jóvenes que actualmente no están realizando ninguna actividad formativa. Sólo se ha podido seleccionar la muestra de jóvenes que viven con sus padres, para poder obtener la información relativa a estos últimos. El modelo se ajustó sobre la variable 'nivel de estudios' restringiendo la respuesta a tres categorías. Tabla 6.1.

Nivel Estudios	Frecuencia	%	Frecuencia	% acumulado
1 (Primaria)	54235	29.96	54235	29.96
2 (Secundaria)	68172	37.66	122407	67.61
3 (Universitarios)	58630	32.39	181037	100

Tabla 6.1 Variable objeto de estudio

A continuación se expresa la probabilidad que tiene cada joven *i* de alcanzar el nivel académico *j*

$$P\{y_i = j|x\} = \frac{e^{(\mathrm{x}'\beta)}}{1+e^{(\mathrm{x}'\beta)}} \qquad (1)$$

$$P\{y_i = \boldsymbol{Estudios\ primarios}|x\} = P\{y^* \leq \mu_1\} = \frac{e^{\mu_1 - \mathrm{x}_i'\beta}}{1 + e^{\mu_1 - \mathrm{x}_i'\beta}}$$

$$P\{y_i = \boldsymbol{Estudios\ secundarios}|x\} = P\{\mu_1 \leq y^* \leq \mu_2\} = \frac{e^{\mu_2 - \mathrm{x}_i'\beta}}{1 + e^{\mu_2 - \mathrm{x}_i'\beta}} - \frac{e^{\mu_1 - \mathrm{x}_i'\beta}}{1 + e^{\mu_1 - \mathrm{x}_i'\beta}}$$

$$P\{y_i = \boldsymbol{Estudios\ secundarios}|x\} = P\{\mu_2 \leq y^*\} = 1 - \frac{e^{\mu_2 - \mathrm{x}_i'\beta}}{1 + e^{\mu_2 - \mathrm{x}_i'\beta}}$$

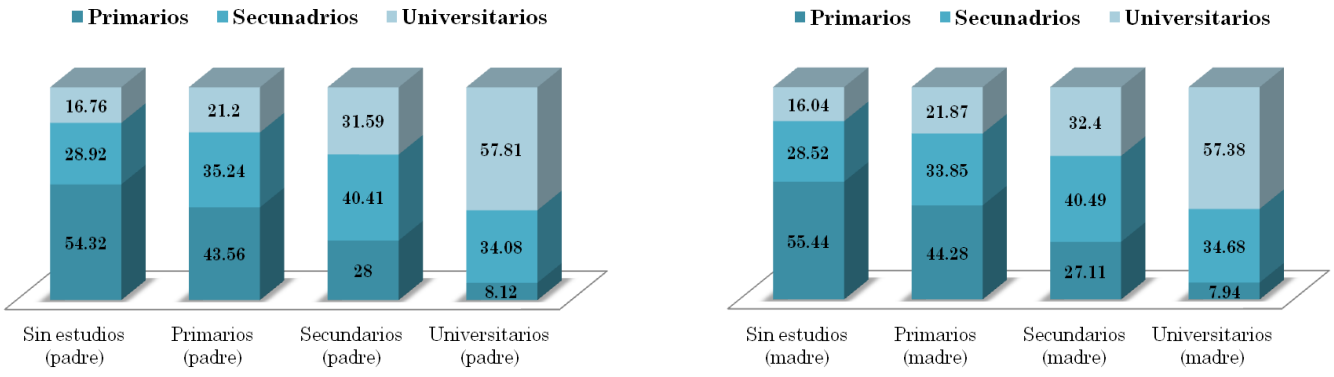


Figura 6.3 Nivel educativo del hijo en función de nivel de los progenitores

En la figura 6.3 se presenta el nivel de formación alcanzado por los hijos según el nivel educativo de sus padres. Se puede apreciar que sendos gráficos presentan distribuciones similares, ya que a medida que aumenta el nivel de estudios de los padres, el nivel alcanzado por los hijos es superior. Cabe destacar que existe relación significativa[1] entre el nivel educativo de los padres y el de los hijos.

Otra de las cuestiones que se han tenido en cuenta en el modelo, es la nacionalidad de los padres, la cual, se presenta en tres categorías como se muestra en la figura 6.4.

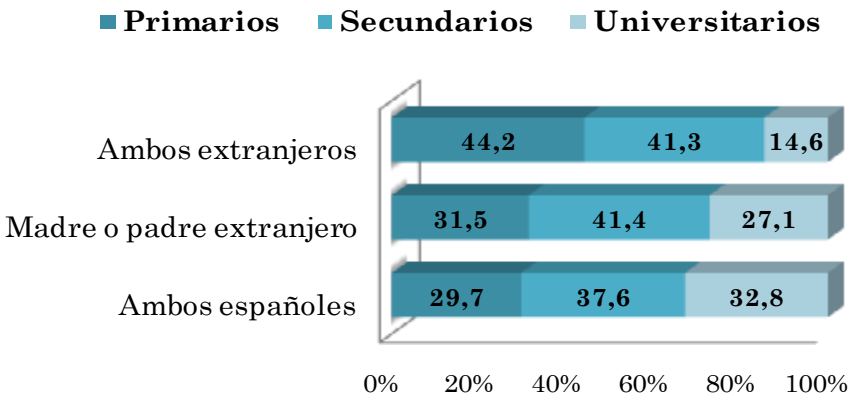


Figura 6.4 Nivel educativo de los hijos según la nacionalidad de los padres

A continuación se presenta la tabla con las variables que se han considerado en el análisis. En ella se pueden agrupar las variables predictoras en dos bloques. En el primero estarían las relacionadas con los aspectos socio-demográficos del hijo. En cuanto al segundo está vinculado con la formación y ocupación de los padres.

Para conocer el comportamiento de las variables que se utilizan en el análisis, se realiza un estudio descriptivo de las mismas (Tabla A3.1, Anexo3).

Variables de modelo logit ordinal		
Sexo	Binaria	
Aréa	Binaria	> 20.000 Hab / < 20.000 Hab
Tenencia vivienda	Binaria	Pagada /Otra situación
Comunidad Autónoma	Nominal	6 grupos de clusters obtenidos en el apartado anetrior
Efecto calendario	Binaria	1ª semestre / 2º semestre
Hermanos	Binaria	No tener hermanos/ Tener 1 o más hermanos
Nacionalidad de los padres	Nominal	Extranjeros/Españoles / padre o madre extranjero/a
Estudios Padre	Nominal	Sin estudios/Primarios/ Secundaria/Universitarios
Estudios Madre	Nominal	Sin estudios/Primarios/ Secundaria/Universitarios
Actividad Padre	Nominal	Empresario/ Antónomo / Trabajador /No aplicable
Actividad Madre	Nominal	Empresario/ Antónomo / Trabajador /No aplicable

Tabla 6.2 Variables consideradas en el modelo

[1] Se ha realizado la prueba X^2 para contrastar la independencia de nivel educativo de los padres frente al de los hijos. Se rechaza la hipótesis nula para cual α razonable (p-valor<.0001 Tabla A3.2-A3.3, Anexo 3)

6.2 Prueba de hipótesis de líneas paralelas

Uno de los supuestos más importantes de este modelo es el de líneas paralelas o *proportional odds* que implica que los parámetros que se estiman para cada variable regresora β, son iguales para todas las categorías de la variable dependiente. Sin embargo, si la hipótesis no se cumple, los estimadores son sesgados e ineficientes. Dicha hipótesis no se cumple en nuestro modelo, dado el p-valor asociado al estadístico chi cuadrado, como se puede comprobar en la siguiente tabla.

Test de puntuación para la suposición de disparidad proporcional		
Chi-cuadrado	**DF**	**Pr > ChiSq**
1371.1251	25	<.0001

Tabla 6.3 Test líneas paralelas

Para saber cuáles son las variables que no cumplen con el supuesto, recurrimos al test de Wald, para contrastar la hipótesis de líneas paralelas de cada una de las variables regresoras, en caso de que alguna de las variables cumpla la condición de líneas paralelas, tendríamos que ajustar un modelo logístico ordinal con proporcionalidad parcial, de lo contrario, se tendría que utilizar como por ejemplo los logit ordenados generalizados.

La prueba de Wald contrasta la hipótesis nula de que los parámetros de regresión β son iguales para todas las respuestas acumuladas. Se estima el modelo con líneas no paralelas, esto quiere que los parámetros β son distintos en el nominador y denominador, y se aplica la prueba de Wald de parámetros iguales. De forma matricial se expresa la hipótesis de líneas paralelas.

$$H_0 : L\beta = c$$

Donde:

$L\ es\ la\ matriz\ de\ coeficientes\ para\ la\ hipótesis\ lineas\ paralelas$

$c\ vector\ \ de\ constantes$

$\beta\ es\ el\ vector\ de\ coeficientes\ de\ regresión$

A la vista de los resultados de la tabla 6.4, se podría concluir que no se rechaza la hipótesis nula planteada en las variables 'Área' y 'Vivienda' para un nivel de significación de 1%.

Resultados del test de la hipótesis lineal			
Variables	**Chi-cuadrado de wald**	**DF**	**Pr > ChiSq**
Nacionalidad padres	2	68.008	0.000
Sexo	1	20.717	0.000
Ciudad	1	1.508	0.219
Vivienda	1	4.991	0.025
Hermanos	1	37.907	0.000
Comunidad Autónoma	5	525.390	0.000
Efecto Calendario	1	11.364	0.001
Estudios padre	3	356.359	0.000
Estudios madre	3	468.744	0.000
Situación padre	3	184.824	0.000
Situación madre	3	114.167	0.000

Tabla 6.4 Resultados prueba Wald

Se podría implementar la prueba de Wald utilizando otros métodos alternativos para evaluar el supuesto de líneas paralelas, entre los que se destaca dos técnicas gráficas, en primer lugar, se grafica los logit acumulativos de las dos variables que cumplen con el supuesto (Figura 6.5). Se puede apreciar en la gráfica de ambas variables que las curvas de logit acumulativo empíricos se mueven de una manera similar mientras se mantiene aproximadamente constante la distancia entre ellos, hecho que apoya el test de Wald, en que los parámetros β son iguales para las distintas respuestas.

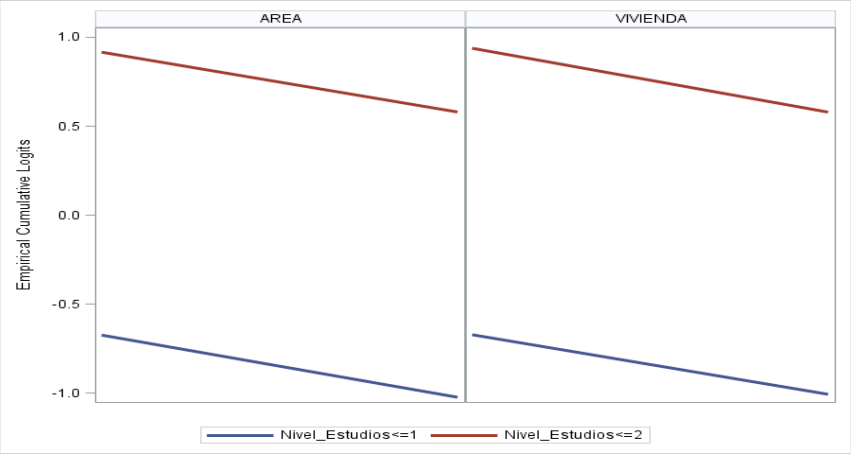


Figura 6.5 Logit Acumulativos

Otra técnica gráfica alternativa que se podría emplear para la evaluación de la proporcionalidad, es la de comparar el valor medio de cada variable independiente dentro de cada nivel de la variable objetivo 'nivel de estudios', con el valor esperado del modelo. Con estos gráficos se podría comprobar dos supuestos:

- Confirmar la ordinalidad de la respuesta para cada regresor, las medias deben ser estrictamente crecientes o decrecientes con respecto a la variable respuesta.
- Evaluar la asunción de proporcionalidad para cada regresor, la curva de valor esperado del modelo debe seguir de cerca la curva media.

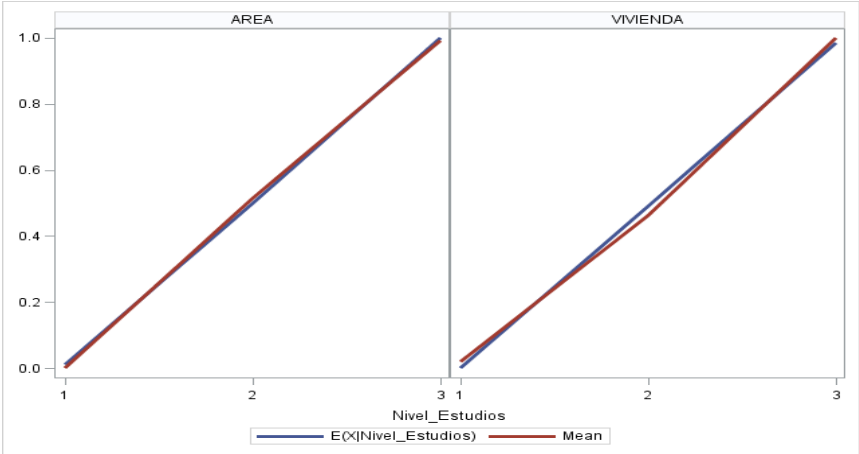


Figura 6.6 Valores medios frente a valores esperados

Para las dos variables en cuestión, se puede comprobar cómo se cumple la ordinalidad con respecto a la variable respuesta, así como el paralelismo que existe entre las curvas del valor esperado y la media. En cambio, cuando esto último no se produce, se obtiene los siguientes resultados, como por ejemplo para las variables 'efecto de calendario' y 'tener hermanos' que no cumplen con el supuesto.

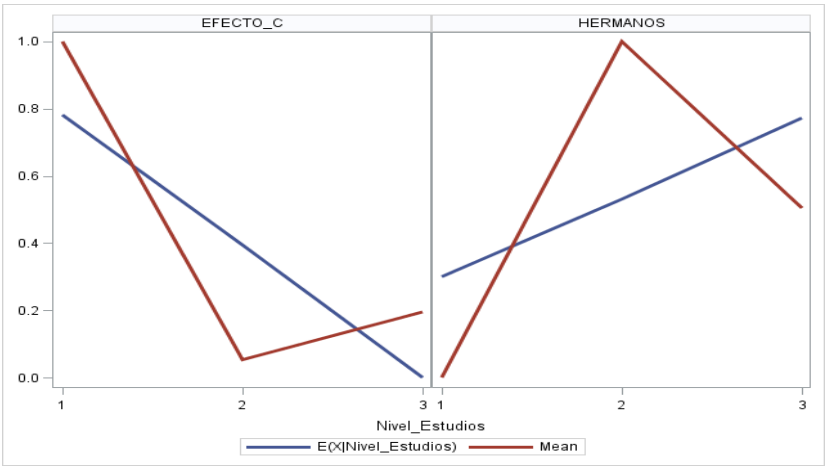


Figura 6.7 Valores medios frente a valores esperados

Las probabilidades de ocurrencia de los valores de la variable dependiente "Nivel de estudios" que se estiman según el modelo para las diferentes combinaciones de las variables 'Efecto calendario' y 'Tener hermanos', divergen significativamente de la frecuencia con la cual ocurren en la muestra los valores de la variable dependiente para estas combinaciones. Lo cual puede ser un claro síntoma de la falta de proporcionalidad.

6.3 Estimación de los parámetros

A continuación, se vuelve ajustar el modelo logístico ordinal generalizado con proporcionalidad parcial, en el cual se especifica que las únicas variables que cumplen con la hipótesis de proporcionalidad son 'Área' y 'Vivienda', lo que se traduce en que

las dos categorías de dichas variables tendrán el mismo parámetro, la única diferencia radicará en el umbral de cada categoría de estudios, en cuando al resto de regresores se estimará un parámetro para cada categoría. Se comprueba la significatividad global de cada variable (Tabla 3.4, Anexo3), en la cual se verifica que todos los regresores considerados en el modelo son significativos para cualquier nivel de significación.

A continuación se muestra la tabla con la estimación de los parámetros de cada variable independiente.

	Análisis del estimador de máxima verosimilitud									
	Categoría	**Referencia**	**Estimador (Primaria)**	**Error estándar**	**Chi cuadrado Wald**	**Pr > ChiSq**	**Estimador (Secuandaria)**	**Error estándar**	**Chi cuadrado Wald**	**Pr > ChiSq**
Constante			0.0693	0.0492	1.9883	0.0985	1.9854	0.0479	1715.1989	<.0001
Sexo	Mujer	Hombre	-0.9611	0.0114	7057.8078	<.0001	-1.0217	0.0109	8837.1962	<.0001
Aréa	> 20.000 Hab	< 20.000 Hab	-0.1433	0.00959	223.2232	<.0001	-0.1433	0.00959	223.2232	<.0001
Tenencia vivienda	Pagada	Otra situación	-0.3215	0.00924	1211.7424	<.0001	-0.3215	0.00924	1211.7424	<.0001
Comunidad Autónoma	Andalucía, Extremadura,Murcia, C. Valenciana, y C. la Mancha	País Vasco, Navarra	0.6685	0.0261	653.6555	<.0001	0.2456	0.0219	125.835	<.0001
	Cantabria, Cataluña, y La Rioja		0.5029	0.0286	308.4055	<.0001	0.1845	0.0243	57.4894	<.0001
	C. León, Asturias y Galicia		0.4923	0.0279	311.4594	<.0001	0.3202	0.0238	180.6832	<.0001
	Aragón y Madrid		0.5292	0.0286	342.8429	<.0001	0.2258	0.0241	88.0546	<.0001
	Canarias y Baleareas		0.6508	0.0356	334.7675	<.0001	0.4292	0.0337	162.5878	<.0001
Efecto calendario	2º semestre	1ª semestre	0.0392	0.0109	12.8472	0.0003	0.0025	0.0107	0.0544	0.8156
Hermanos	Tener 1 o más hermanos	0 hermanos	0.0352	0.0116	9.2273	0.0024	0.0914	0.0114	64.0911	<.0001
Nacionalidad padres	Alguno extranjero	Española	0.2582	0.0519	24.7289	<.0001	0.456	0.0537	72.1047	<.0001
	Ambos extranjeros		0.4561	0.04	130.3309	<.0001	0.9584	0.0538	317.9035	<.0001
Estudios Padre	Primarios	Sin estudios	-0.1869	0.0236	62.9658	<.0001	-0.1165	0.0296	15.5005	<.0001
	Secundaria		-0.6068	0.0227	717.1371	<.0001	-0.491	0.028	307.9774	<.0001
	Universitarios		-1.677	0.033	2581.0158	<.0001	-1.3118	0.0315	1736.6871	<.0001
Estudios Madre	Primarios	Sin estudios	-0.255	0.0242	111.0034	<.0001	-0.276	0.0308	80.4424	<.0001
	Secundaria		-0.7726	0.0237	1065.107	<.0001	-0.6252	0.0296	444.7863	<.0001
	Universitarios		-1.7092	0.0355	2313.5201	<.0001	-1.2735	0.0337	1425.2171	<.0001
Actividad Padre	Antónomo	Empresario	0.0401	0.0253	2.5059	0.1134	0.1081	0.0235	21.1927	<.0001
	Trabajador		0.2106	0.0208	102.0492	<.0001	0.2922	0.0191	235.2084	<.0001
	No aplicable		0.1586	0.0223	50.5892	<.0001	0.1305	0.0207	39.6848	<.0001
Actividad Madre	Antónoma	Empresaria	0.0525	0.0385	1.8632	0.1723	0.039	0.0347	1.2648	0.2608
	Trabajador		0.1992	0.0331	36.2642	<.0001	0.172	0.0293	34.3953	<.0001
	No aplicable		0.2102	0.0334	39.6537	<.0001	0.1321	0.0299	19.5902	<.0001

Tabla 6.5 Estimación de los parámetros

En la tabla 6.5 se expone las estimación de los parámetros, en cual se debe de tener en cuenta que la columna 'Estimador primaria' muestra el nivel educativo primaria comparado con los niveles educativos secundaria y universitario, y en cuanto a la columna de 'Estimador Secundaria' muestra el nivel educativo primario y secundario frente a los estudios universitarios. Así como el correspondiente error estándar y el p-valor asociado a cada uno de los parámetros.

La interpretación de estos estimadores se lleva a cabo de la misma manera que en los estimadores de la regresión logística binaria. Para cuantificar estas probabilidades recurrimos a los odds radio.

6.4 Interpretación del modelo

Como se ha comentado en el anterior apartado las comparaciones del modelo logit ordenado acumulado, se realizarán de la siguiente manera:

Comparación 1: "Estudios primarios" (ref) frente a "Estudios secundarios" y "Estudios universitarios" (columna Odds ratio primaria de la tabla 6.6).

$$Odds\ ratio = \frac{P\{y\ \leq Estudios\ primarios|x = x_2\}/P\{y > Estudios\ primarios|x = x_2\}}{P\{y\ \leq Estudios\ primarios|x = x_1\}/P\{y > Estudios\ primarios|x = x_1\}}$$

Comparación 2: "Estudios primarios" y "Estudios secundarios" (ref) frente a "Estudios universitarios" (columna Odds ratio secundaria de la tabla 6.6).

$$dds\ ratio = \frac{P\{y\ \leq Estudios\ secundarios|x = x_2\}/P\{y > Estudios\ secundarios|x = x_2\}}{P\{y\ \leq Estudios\ secundarios|x = x_1\}/P\{y > Estudios\ secundarios|x = x_1\}}$$

Si el valor de Odds Ratio es menor que uno, lo cual sucede cuando el coeficiente de la variable regresora es negativo, indica que, si las otras variables explicativas permanecen constantes, los cambios en la variable explicativa analizada incrementan la probabilidad de obtener categorías de mayor valor en la variable objeto de estudio. En cambio, si los valores de Odds Ratio son mayores que uno, esto demuestra que las variaciones en la variable independiente disminuye el riesgo de obtener categorías de mayor valor de la variable objetivo. A continuación se procede a interpretar los Odds Ratio más destacados. Las categorías señaladas en rojo no presentan diferencias significativas respecto a la categoría de referencia.

	Categoría	**Referencia**	**Odds Ratio (Primarios)**	**Límites de confianza al 95% de Wald**		**Odds Ratio (Secundarios)**	**Límites de confianza al 95% de Wald**	
Sexo	Mujer	Hombre	**0.382**	**0.374**	**0.391**	**0.36**	**0.352**	**0.368**
Aréa	> 20.000 Hab	< 20.000 Hab	0.866	0.85	0.883	0.866	0.85	0.883
Tenencia vivienda	Pagada	Otra situación	0.725	0.712	0.738	0.725	0.712	0.738
Comunidad Autónoma	Andalucía, Extremadura,Murcia, C. Valenciana, y C. la Mancha	País Vasco, Navarra	**1.951**	**1.854**	**2.054**	**1.278**	**1.225**	**1.334**
	Cantabria, Cataluña, y La Rioja		1.653	1.563	1.749	1.203	1.147	1.261
	C. León, Asturias y Galicia		1.636	1.549	1.728	1.377	1.315	1.443
	Aragón y Madrid		1.698	1.605	1.795	1.253	1.196	1.314
	Canarias y Baleareas		**1.917**	**1.788**	**2.056**	**1.536**	**1.438**	**1.641**
Efecto calendario	2º semestre	1ª semestre	1.04	1.018	1.063	1.003	0.982	1.024
Hermanos	Tener 1 o más hermanos	0 hermanos	1.036	1.013	1.06	1.096	1.071	1.121
Nacionalidad padres	Alguno extranjero	Españoles	1.295	1.169	1.433	1.578	1.42	1.753
	Ambos extranjeros		**1.578**	**1.459**	**1.707**	**2.608**	**2.347**	**2.897**
Estudios Padre	Primarios	Sin estudios	0.83	0.792	0.869	0.89	0.84	0.943
	Secundaria		0.545	0.521	0.57	0.612	0.579	0.646
	Universitarios		**0.187**	**0.175**	**0.199**	**0.269**	**0.253**	**0.286**
Estudios Madre	Primarios	Sin estudios	0.775	0.739	0.813	0.759	0.714	0.806
	Secundaria		0.462	0.441	0.484	0.535	0.505	0.567
	Universitarios		**0.181**	**0.169**	**0.194**	**0.28**	**0.262**	**0.299**
Actividad Padre	Antónomo	Empresario	1.041	0.99	1.094	1.114	1.064	1.167
	Trabajador		**1.234**	**1.185**	**1.286**	**1.339**	**1.29**	**1.39**
	No aplicable		1.172	1.122	1.224	1.139	1.094	1.187
Actividad Madre	Antónoma	Empresaria	1.054	0.977	1.136	1.04	0.971	1.113
	Trabajador		**1.22**	**1.144**	**1.302**	**1.188**	**1.121**	**1.258**
	No aplicable		1.234	1.156	1.317	1.141	1.076	1.21

Tabla 6.6 Odds Ratio modelo logístico ordinal

- En el caso de la variable sexo se tiene que, el riesgo de las jóvenes de alcanzar estudios secundarios o superiores es mayor en comparación con los jóvenes varones, *ceteris paribus*.

- Lo mismo ocurre cuando se compra la categoría de estudios primarios o secundarios frente a los estudios universitarios, las mujeres siguen teniendo mayor riesgo de conseguir estudios universitarios, *ceteris paribus*.
- En cuanto al nivel de estudios de los padres (se obtienen resultados similares en lo que respecta a las madres y los padres), se ve como el riesgo de lograr estudios secundarios o universitarios es 5,5 menor cuando los padres no tienen estudios frente a los jóvenes que sus padres tienen estudios universitarios. En cuanto al riesgo de alcanzar los estudios universitarios para los jóvenes cuyos padres no tienen estudios es 3.5 veces menor frente a los jóvenes que sus padres tienen estudios universitarios, *ceteris paribus*.
- En lo referente a la nacionalidad de los padres, se obtiene que los jóvenes cuyos padres son españoles tienen un riesgo mayor de alcanzar estudios universitarios frente a los jóvenes de padres extranjeros, *ceteris paribus*.
- El riesgo de alcanzar estudios universitarios entre los jóvenes andaluces, valencianos, murcianos, extremeños y manchegos disminuye en aproximadamente dos veces frente a los jóvenes vascos y navarros, *ceteris paribus*.
- Por último, los jóvenes de padres empresarios tienen mayor de riesgo de alcanzar estudios universitarios frente a los jóvenes cuyos padres son trabajadores por cuenta ajena, *ceteris paribus*.

En lo que respecta a la bondad de ajuste de este modelo, las pruebas revelan que no hay evidencia para considerar que hay falta de ajuste

7 Ocupación laboral de los jóvenes universitarios

En este apartado se va a centrar el foco en los jóvenes con empleo que tienen estudios universitarios, y mediante el análisis de correspondencias simple 3.3, se va a investigar con que ocupaciones se relaciona cada tipo de estudios.

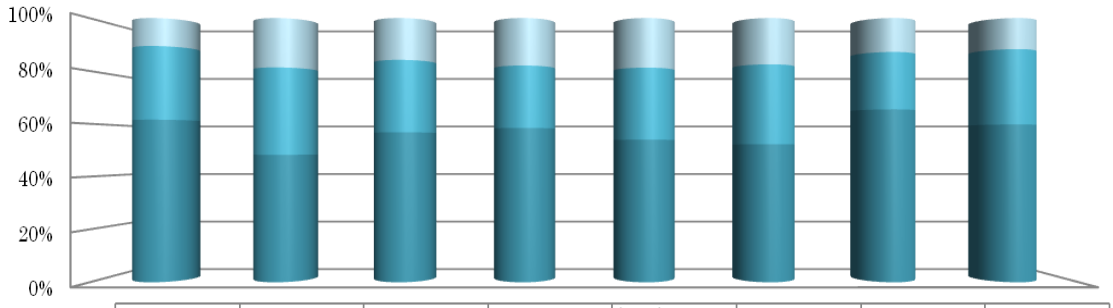
En la fuente de datos se tiene la información sobre la clasificación nacional de ocupaciones (CNO), estas clasificaciones son estructuras elaboradas con el objeto de poder agrupar unidades homogéneas, según un criterio definido, en una misma categoría. La información que se agrupa en dicha variable se recoge en la siguiente tabla.

Categoria	Descripción
0	Ocupaciones militares
1	Directores y gerentes
2	Técnicos y profesionales científicos e intelectuales
3	Técnicos; profesionales de apoyo
4	Empleados contables, administrativos y otros empleados de oficina
5	Trabajadores de los servicios de restauración, personales, protección y vendedores
6	Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero
7	Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción (excepto operadores de instalaciones y maquinaria)
8	Operadores de instalaciones y maquinaria, y montadores
9	Ocupaciones elementales

Tabla 7.1 Clasificación nacional de ocupaciones (CNO-2011)

Como se muestra en la tabla, la clasificación nacional de ocupaciones se distribuye en diez categorías, mediante estas vamos a identificar que carreras universitarias se relacionan con cada una de ellas. Cabe mencionar que cada categoría se puede desagregar en varias subcategorias[2]. En nuestro caso, se ha decidido trabajar con las categorías de la tabla 7.1 para una mayor simplicidad de los resultados, ya que la categoría desagregada consta de 90 niveles.

Antes de empezar el análisis, se muestra la proporción de ocupados y parados por tipo de estudios. Cabe destacar de la Figura 7.1, que los jóvenes que han estudios las carreras de la rama de artes y humanidades presenta la mayor tasa de parados con respecto al resto. En cambio, los jóvenes con estudios en áreas de salud tienen la menor tasa de desempleo y la mayor tasa de ocupación.



	Educación	Artes y Humanidades	Derecho y Ciencias Sociales	Ciencias	Arquitectura, Construcción, Formación Técnica e Industrias	Agricultura, Ganadería, Pesca y Veterinaria	Salud y Servicios Sociales	Otros servicios
Otra situación	10.57	18.68	15.83	17.96	18.77	17.52	12.84	11.76
Paro	27.89	33.06	27.4	23.61	27.27	30.26	21.72	28.51
Ocupado	61.54	48.25	56.77	58.43	53.96	52.22	65.44	59.73

Figura 7.1 Tabla de contingencias de actividad laboral por tipo de estudios

En las siguientes tablas, vemos que aproximadamente uno de cada cuatro jóvenes trabaja a tiempo parcial, de los cuales el 72% son mujeres. Esto último podría atribuirse a la maternidad, ya que el 92% de estas mujeres son madres.

[2] Se puede consultar la desagregación en el siguiente enlace:
http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t40/cno11&file=inebase

Jornada de trabajo	Frecuencia	Porcentaje
Tiempo completo	46723	76.86
Tiempo parcial	14066	23.14

Sexo	Frecuencia	Porcentaje
Hombre	3997	28.42
Mujer	10069	71.58

Tabla 7.2 Jornada de trabajo y Distribución sexo de los jóvenes a tiempo parcial

7.1 Análisis de correspondencias simples

Se recurre al análisis de correspondencias simples. Es una técnica que a partir de la tabla de contingencia o correspondencias de dos variables cualitativas permite determinar sí existen asociaciones entre sus categorías (Apartado 3.3).

Estadístico	DF	Valor	Prob
Chi-cuadrado	56	17566.0667	<.0001
Chi-cuadrado de ratio de verosimilitud	56	16771.441	<.0001
Chi-cuadrado Mantel-Haenszel	1	265.9763	<.0001
Coeficiente Phi		0.5376	
Coeficiente de contingencia		0.4735	
V de Cramer		0.2032	

Tabla 7.3 Contraste Chi Cuadrado

Un paso previo consiste en comprobar si existe asociación entre las variables objeto de estudio, para ello calculamos el estadístico X^2, en base a las diferencias entre los valores observados y esperados, y de su análisis se desprende que existen diferencias importantes. Por esta razón la X^2 muestra un valor estadísticamente significativo, p-valor<.0001.

Inercia y descomposición chi-cuadrado					
Valor singular	**Inercia principal**	**Chi-cuadrado**	**Porcentaje**	**Porcentaje acumulado**	14 28 42 56 70 ----+----+----+----+----+---
0.44435	0.19744	12002.4	68.33	68.33	**********
0.24919	0.0621	3774.8	21.49	89.82	********
0.12311	0.01516	921.3	5.24	95.06	**
0.11137	0.0124	754	4.29	99.35	**
0.03365	0.00113	68.8	0.39	99.75	
0.02522	0.00064	38.7	0.22	99.97	
0.01004	0.0001	6.1	0.03	100	
Total	0.28897	17566.1	100		
Grados de libertad = 56					

Tabla 7.4 Descomposición de la inercia

Se observa que la inercia total vale 0.2889, y el valor del estadístico chi-cuadrado X^2 = N(λ_1+ … + λ_7) =17566.1. De los resultados se determina que la hipótesis de independencia es rechazada ya que 17566.1 > $X^2_{56}(\alpha)$ para cualquier α razonable.

7.2 Determinación de número de dimensiones

En cuanto a la elección de factores a retener, se ha basado en la aplicación del criterio que consiste en tomar un número de factores que expliquen un porcentaje suficiente de

la información, por lo que se retiene los dos primeros factores dado que explican un 89.82%. Elegido el número de factores vamos a describirlos utilizando para ello las contribuciones parciales, en el anexo (Tabla A4.1, Anexo4) se presenta la información con todos los estadísticos relativos al análisis.

La contribución absoluta muestra la contribución de las modalidades a la formación del eje. Si analizamos las categorías de la variable tipo de estudios que mayor contribución tienen en la formación del primer eje cabe destacar 'Derecho y Ciencias Sociales' y 'Otros servicios', las dos categorías contribuyen conjuntamente con más de un 60%, y la categoría 'Salud y Servicios Sociales' con un 24%. En conjunto explican más de 84% de la información que contiene el primer factor. En cuanto al segundo factor, las categorías que más aportan a la formación de este último son las categorías 'Ciencias', 'Arquitectura, Construcción, Formación Técnica e Industrias' y 'Salud y Servicios Sociales', con 32%, 28% y 24% respectivamente. Con estas tres categorías se explica más de 85%.

Se repite el análisis con las modalidades columna de la variable 'Ocupación', para la formación cabe resaltar la categoría 'Empleados contables, administrativos y otros empleados de oficina' que contribuye de forma muy importante con 60%, seguido de la categoría 'Técnicos y profesionales científicos e intelectuales' con un 31%. En total aportan más de 91% a la creación del primer eje. En cuanto al segundo eje, se tiene que la categoría 'Técnicos; profesionales de apoyo' aporta un 75% a la formación del segundo eje.

En cuanto a la calidad, se tiene que la comunalidad de todas las categorías es admisible, a excepción de 'Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero', cabe apuntar que las dos dimensiones consiguen explicar un porcentaje alto de la variabilidad de todas las categorías.

7.3 Interpretación del gráfico de representación conjunta

Finalmente se representa un gráfico conjunto todas las categorías de ambas variables, en los mismos ejes dimensionales (Figura 7.2).

Desde un punto de vista gráfico, se analiza las relaciones de conjunto de las dos variables categóricas. En el mapa simétrico se vuelven a poner de manifiesto las relaciones de dependencia existentes entre las dos variables.

- Se puede apreciar cómo los estudios de derecho y ciencias sociales (Administración y dirección de empresas, Psicología, Economía, Periodismo, entre otras) se encuentran asociadas con los cargos de directores y gerentes.
- En lo que respecta a las carreras en el área de artes y humanidades (Historia, Lenguas, Imagen y sonido, entre otras), vemos como estos titulados se encuentran relacionados con las ocupaciones elementales y el sector de los servicios de restauración.

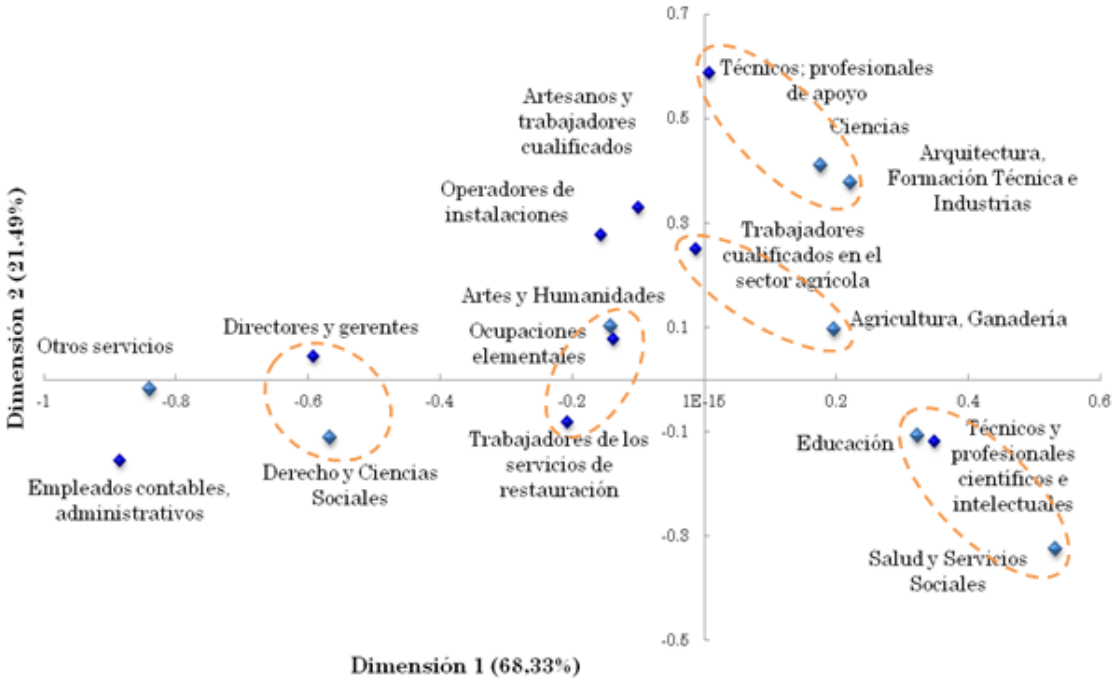


Figura 7.2 Análisis de correspondencias de Tipo de estudios por ocupación

- En cuanto a la ramas de salud y servicios sociales (Medicina, enfermería, farmacia, etc) y Educación, estas especialidades se vinculan con la categoría de Técnicos y profesionales científicos e intelectuales.
- Con los cargos de técnicos y profesionales de apoyo, estarían asociados los jóvenes con estudios en las áreas de ciencias (Biología, Química, Física, Matemáticas, etc) e informática, y arquitectura, formación técnica e industrias.
- Por último, los trabajadores cualificados en el sector agrícola, son jóvenes con preparación en dichas áreas (veterinaria, ingeniería agrónoma o similar).

8 Factores que influyen en la emancipación

La aspiración a la emancipación es compartida por todos los jóvenes europeos, pero este hecho se realiza de diferentes maneras y en diferente tiempo. A continuación se presenta

en la figura 8.1 la información relativa a la tasa de jóvenes de 22 a 29 años que viven con sus padres.

A la vista de los datos, se puede apreciar que España presenta una tasa de emancipación de 32.1%, doce puntos porcentuales por debajo de la media de la Unión Europea. Este ranking lo encabeza los países nórdicos donde el camino a la autonomía se presenta a una edad temprana (20 años) entre la formación y el periodo de empleo, estos países tienen un sistema de apoyo entre subvenciones y préstamos que facilitan la independencia.

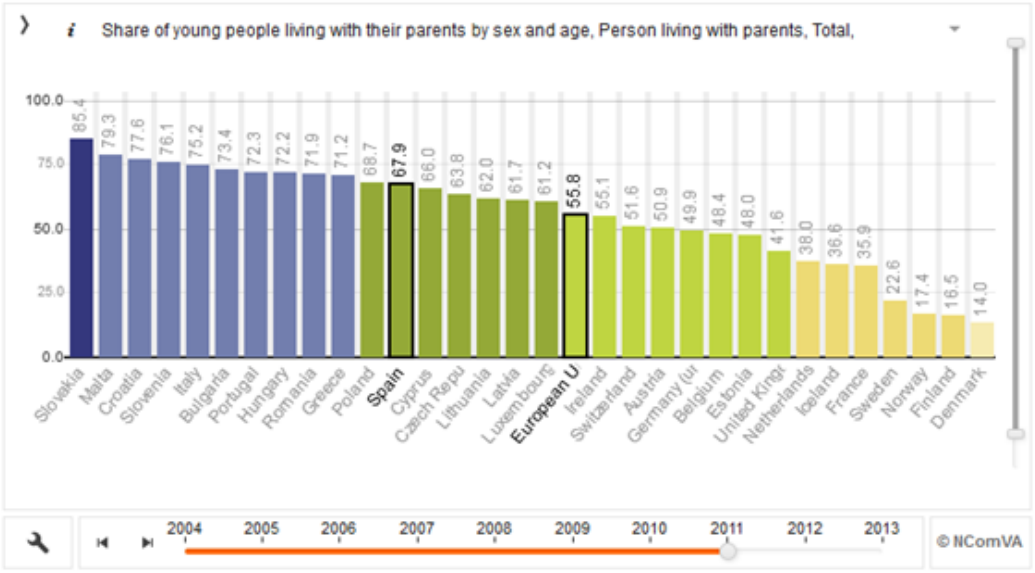


Figura 8.1 Proporción de jóvenes que viven con sus padres (fuente Eurostat)

En la otra cola nos encontramos a la mayoría de los países del mediterráneo, en los que la emancipación apenas alcanza el 30% entre los jóvenes de 22 a 29 años, en estos países predomina ciertos factores a la hora de tomar la decisión de abandonar el hogar familiar, entre los que se destaca, los niveles salariales y la estabilidad en el empleo, ya que estás juegas un papel importante a la hora de acceder al régimen de tenencia en propiedad (Juan Manuel Patón Casas) [22] .

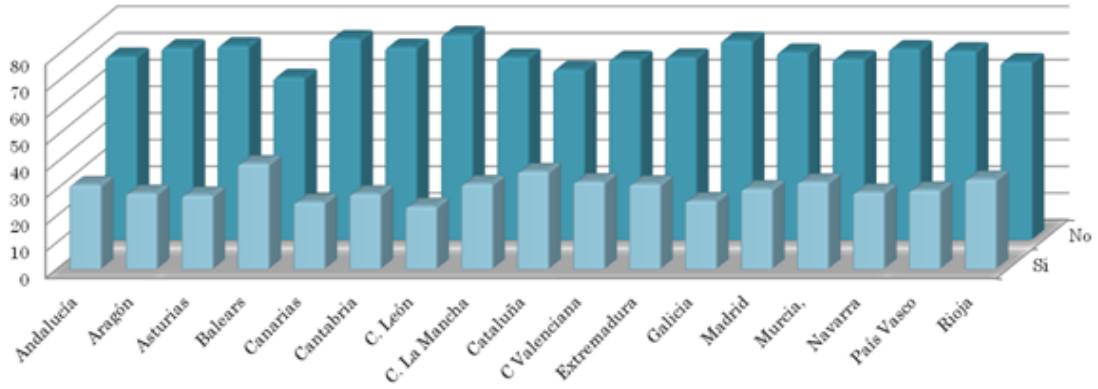


Figura 8.2 Tasa de emancipación por comunidad autónoma

Volviendo de nuevo a nuestros datos, vemos que la mayor tasa de emancipación se presenta en las Islas Baleares rozando casi el 40%, casi diez puntos por encima de la media nacional, así lo constato el instituto de política familiar de Baleares en 2011.

Una vez analizados los resultados de la muestra a nivel de comunidades autónomas, nos vamos adentrar en los aspectos socio-demográficos que atañen a los jóvenes de la muestra. En la figura 8.3 se observa que la proporción de jóvenes menores de 30 años emancipadas es 57.88% frente al 42.12% de los hombres de su misma edad, existiendo diferencias significativas entre los sexos[3]. Estas diferencias se observan en las distintas edades que se han estudiado (22 a 30 años).

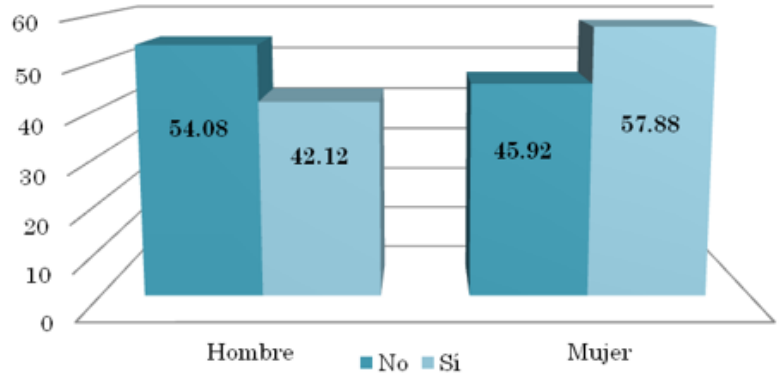


Figura 8.3 Situación de emancipación por sexo

Si analizamos solo los jóvenes que tienen empleo, vemos que estas diferencias se manifiestan del mismo modo entre los dos sexos.

Área de residencia	Emancipado		Total
	No	Sí	
Municipio ≤ 20.000 hab	120713 69.47	53056 30.53	173769
Municipio > 20.000 hab	139206 69.29	61707 30.71	200913
Total	259919	114763	374682

8.1 Tabla contingencia área de residencia por emancipación

En cuanto al área de residencia, no se aprecian diferencias significativas en cuanto a la tasa de emancipación.

8.1 Construcción del modelo logístico binario múltiple

La técnica estadística empleada en esta sección, es la Regresión Logística binaria, que permite modelar el comportamiento de una variable categórica, es decir, que mide la presencia de una característica del individuo, en función de una serie de variables predictivas que podrán ser de naturaleza continua o categórica.

[3] Se utilizó la prueba X^2 de Pearson para contrastar la igualdad de proporciones

El resultado de esta técnica es la construcción de una función que predice el valor de la probabilidad de que la variable respuesta tome el valor de referencia para cada combinación de valores de las variables significativas, es decir, las que tienen capacidad de predicción sobre la variable respuesta. Se conocerá, así mismo, la influencia que tiene cada variable predictiva sobre la respuesta y también, y a través de los OR (Odds Ratio), el cambio en la probabilidad estimada de pertenencia al valor modelado de la variable respuesta que supone el cambio de categoría de cada variable del modelo.

Modelo de regresión logística

Variable independiente situación de emancipación $= \begin{cases} 0, & \text{No} \\ 1, & \text{Sí} \end{cases}$

Las variables independientes son las presentadas en la tabla 8.2 y se representan por: $X' = (X_1,\ldots,X_p)$, con $x=(x_1,\ldots, x_p)$.

$\pi(x) = P\{Y = 1 \ / \ X = x\} = \frac{e^{g(x)}}{1+ e^{g(x)}}$, donde $g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ es el logit del modelo.

A continuación se presenta la tabla con las variables que se han considerado en el análisis. El análisis descriptivo de las mismas es el presentado en el aparto de la fase de exploración y modificación de SEMMA, y en la tabla A5.1, Anexo 5.

Variable	Descripción	Categorías
Nacionalidad	Binaria	
Sexo	Binaria	
Estado Civil	Nominal	Soltero/ Casado / Separado
Estudios	Nominal	Primarios/ Secundarios /Bachiller / FP / Universitarios
Actividad	Nominal	Ocupado / Paro / Otra situación
Edad	Continua	
Área de residencia	Binaria	> 20.000 Hab / < 20.000 Hab
CCAA	Nominal	17 comunidades autónomas

Tabla 8.2 Variables que intervienen en la construcción del modelo

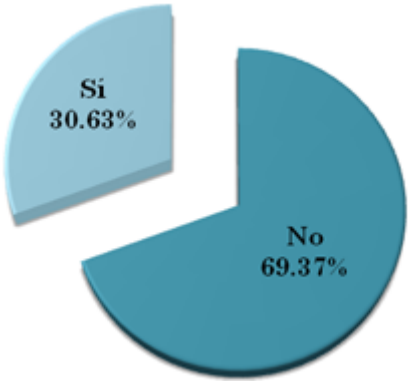


Figura 8.4 Distribución de la situación de emancipación

La variable dicotómica situación de emancipación es la variable objetivo para el análisis logístico. En nuestro estudio el 69,37% de los jóvenes de entre 22 y 30 años viven con sus padres, frente al 30,63% de los jóvenes que si se encuentran emancipados.

Para comenzar con el procedimiento logístico, lo primero será decidir cuáles de las variables independientes del cuestionario relacionadas con los jóvenes son significativas en el modelo. Para ello se utiliza el procedimiento *Stepwise*, basado en la combinación de *forward* y *backward*. En el primer paso se procede como en el método forward, pero a diferencia de éste en el que cuando una variable entra en el modelo ya no vuelve a salir, en el procedimiento *Stepwise* es posible que la inclusión de una nueva variable haga que otra que ya estaba en el modelo resulte redundante y sea "expulsada" de él. Antes de la elección de una nueva variable a incluir, comprueba que todas las seleccionadas con anterioridad siguen siendo significativas, utilizando así, la técnica del procedimiento *backward*. (Significatividad global variables tabla A5.2, Anexo 5).

8.2 Estudio de las posibles interacciones

Como norma general, primero deben ajustarse los modelos de regresión logística con las variables independientes, este modelo se le denomina modelo de efectos principales. Posteriormente, se procede a estudiar las posibles interacciones existentes entre las variables, el modelo presenta ocho variables explicativas, lo que quiere decir, que hemos analizado mediante el test de razón de verosimilitud, la significación de 28 posibles interacciones $\binom{8}{2}$. Nuestro modelo se mantiene con los efectos fijos antes mencionados, dado que ninguna interacción resulto ser significativa.

Análisis del estimador de máxima verosimilitud						
Efecto	**Niveles**	**DF**	**Estimador**	**Error estándar**	**Chi-cuadrado de Wald**	**Pr > ChiSq**
Intercept		1	-7.0067	0.0504	19352.0297	<.0001
Nacionalidad (ref= Española)	Extranjera	1	0.514	0.00744	4776.579	<.0001
Sexo (ref=Hombre)	Mujer	1	0.2165	0.00435	2473.6245	<.0001
EstadoCivil (ref=Separado/Divorciado)	Soltero	1	-0.9364	0.0116	6533.591	<.0001
	Casado	1	1.2157	0.0129	8912.4016	<.0001
Estudios (ref= Primarios)	Secundaria	1	0.1924	0.0078	608.2282	<.0001
	Bachiller	1	-0.1066	0.0104	105.1883	<.0001
	Fp	1	0.1179	0.00844	194.9915	<.0001
	Universitarios	1	-0.2942	0.00816	1299.9456	<.0001
Actividad (ref=ocupado)	Paro	1	-0.0234	0.00708	10.9181	0.001
	Otra Situación	1	-0.4147	0.00935	1965.973	<.0001
Edad		1	0.2628	0.00181	21077.7592	<.0001
Área (ref=Municipio < 20.000 hab)	Municipio > 20.000 hab	1	0.0235	0.00465	25.5473	<.0001
CCAA (ref= C. Madrid)	Andalucía	1	0.1128	0.0112	101.0917	<.0001
	Aragón	1	-0.1268	0.0218	33.8632	<.0001
	Asturias	1	-0.1179	0.0313	14.174	0.0002
	Balears	1	0.3544	0.0294	145.6818	<.0001
	Canarias	1	-0.2101	0.0233	81.3239	<.0001
	Cantabria	1	-0.1344	0.036	13.9343	0.0002
	C. León	1	-0.2861	0.0157	334.1807	<.0001
	C. La Mancha	1	0.0634	0.0168	14.2155	0.0002
	Cataluña	1	0.2535	0.0117	465.6852	<.0001
	C Valenciana	1	0.1166	0.0142	67.5176	<.0001
	Extremadura	1	0.2726	0.0214	162.2831	<.0001
	Galicia	1	-0.2624	0.0193	185.1602	<.0001
	Murcia	1	-0.00537	0.0241	0.0495	0.8239
	Navarra	1	-0.1289	0.0296	18.9679	<.0001
	País Vasco	1	0.00523	0.0195	0.072	0.7885
	Rioja	1	0.1203	0.0422	8.1127	0.0044

Tabla 8.3 valores estimados para cada categoría

La tabla 8.4 recoge los valores estimados para los coeficientes del modelo, junto con sus p-valores asociados. Los coeficientes βi, se interpretan como el cambio que se produce en el término Logit al incrementarse en una unidad la variable explicativa asociada. En el siguiente apartado se interpretan los odds-ratio.

8.3 Interpretación del modelo

En la tabla 8.4 se recogen los OR estimados para cada variable. Los odds ratio señalados en rojo no presentan diferencias estadísticamente significativas respecto a la categoría referencia.

- ***Estado civil:*** Variable categórica siendo la referencia estar separado o divorciado. Se tiene que el riesgo de estar emancipado es de 2 veces mayor cuando se está separado o divorciado frente a los solteros, en cambio el riesgo de estar emancipados entre los casados es 4.5 mayor respecto a los solteros, *ceteris paribus*.

Efecto	**Niveles**	**Estimador del punto**	**Límites de confianza al 95% de Wald**	
Nacionalidad (ref= Española)	Extranjera	2.795	2.715	2.878
Sexo (ref=Hombre)	Mujer	1.542	1.516	1.568
EstadoCivil (ref=Separado/Divorciado)	Soltero	0.518	0.487	0.551
	Casado	4.459	4.181	4.755
Estudios (ref= Primarios)	Secundaria	1.107	1.074	1.141
	Bachiller	0.821	0.793	0.85
	Fp	1.028	0.995	1.061
	Universitarios	0.681	0.659	0.703
Actividad (ref=ocupado)	Paro	0.63	0.618	0.643
	Otra Situación	0.426	0.414	0.438
Edad		1.301	1.296	1.305
Área (ref=Municipio < 20.000 hab)	Municipio > 20.000 hab	1.048	1.029	1.067
CCAA (ref= C. Madrid)	Andalucía	1.15	1.115	1.186
	Aragón	0.905	0.86	0.952
	Asturias	0.913	0.853	0.978
	Balears	1.464	1.373	1.561
	Canarias	0.833	0.79	0.877
	Cantabria	0.898	0.831	0.971
	C. León	0.772	0.742	0.803
	C. La Mancha	1.095	1.05	1.141
	Cataluña	1.324	1.282	1.366
	C Valenciana	1.154	1.114	1.196
	Extremadura	1.349	1.283	1.418
	Galicia	0.79	0.755	0.827
	Murcia	1.022	0.968	1.078
	Navarra	0.903	0.846	0.964
	País Vasco	1.033	0.987	1.081
	Rioja	1.159	1.058	1.269

Tabla 8.4 Estimación de los OR

- ***Área de residencia:*** Se observa que los jóvenes residentes en municipios con menos de 20.000 habitantes tienen mayor riesgo de emanciparse frente a los que viven en municipios con más habitantes, *ceteris paribus*
- ***Nacionalidad:*** Variable dicotómica. Categoría de referencia nacionalidad española. OR=2.79 con $IC_{95\%}$=(2.715; 2.878), el riesgo de emancipación es 2.7 veces mayor si el individuo tiene nacionalidad extranjera frente a los jóvenes de nacionalidad española, *ceteris paribus.*
- **Sexo:** Variable dicotómica. Categoría de referencia hombre. OR=1.54 con $IC_{95\%}$=(1.516; 1.568), vemos que las jóvenes tienen mayor riesgo de emancipación frente a los hombres, esta conclusión corrobora el resultado obtenido en el análisis descriptivo, *ceteris paribus.*
- ***Edad:*** En el caso de la edad, el OR es de 1.301. El intervalo de confianza hallado presenta un límite inferior de 1.295 y un límite superior de 1.305. Se multiplica por 1.3 el riesgo de emanciparse, en el caso de 5 años, se multiplicaría el riesgo por 3.7. Es de esperar que el riesgo de emancipación crezca de forma muy significativa con el aumento de la edad, ya que como se ha comentado el proceso de emancipación se produce en los últimos años del intervalo estudiado, *ceteris paribus.*
- ***CCAA***: La comunidad autónoma que mayor riesgo presenta de emanciparse respecto a la comunidad de Madrid son las Islas Baleares, hecho que se ha probado en el análisis descriptivo que se ha realizado al principio de este epígrafe, seguidas de Cataluña y Extremadura. En cambio, los jóvenes de la comunidad de Madrid tienen mayor riesgo respecto a los jóvenes gallegos y manchegos. En cuando a los jóvenes vascos y murcianos no presentan diferencias significativas respecto a los madrileños, *ceteris paribus.*
- ***Estudios:*** En la población en estudio, el riesgo de emanciparse teniendo estudios primarios es de 1.4 veces mayor respecto a los jóvenes con estudios universitarios, resultado que se debe a que estos últimos se incorporan tarde al mercado laboral debido a la trayectoria académica, *ceteris paribus.*
- ***Actividad:*** Variable categórica siendo la referencia estar ocupado, el riesgo de emanciparse es 2.3 veces mayor cuando se tiene empleo frente a los jóvenes que se encuentran en otra situación (rentistas, personas con invalidez). En cuanto a las personas que están en situación de desempleo, el riesgo de emanciparse disminuye frente a los que tiene empleo, *ceteris paribus.*

8.4 Evaluación de la idoneidad del modelo

En este apartado evaluamos la idoneidad del modelo, para ello existen varias medidas de bondad de ajuste, que se presentan a continuación.

- **Tabla de clasificación**

La tabla de clasificación es normalmente el criterio que debemos de seguir para indicar la bondad de ajuste del modelo.

Tabla de clasificación					
Nivel de prob	**Porcentajes**				
	Correcto	**Sensibilidad**	**Especificidad**	**Falso Positivo**	**Falso Negativo**
0.5	90.5	74.3	95.6	15.4	7.9

Tabla 8.5 Tabla de clasificación

Como se puede observar en la tabla 8.6, el porcentaje de clasificación correcta, es superior en el grupo 0 que el 1, el problema que se nos presenta es el distinto tamaño de cada grupo (70% - 30%), ya que la clasificación siempre favorece el grupo más numeroso. Por lo tanto debemos determinar el punto de corte adecuado, ya que los resultados que están basados en la sensibilidad y especificidad, dependen de la composición de la muestra. Por lo tanto, para evaluar la bondad de nuestro análisis debemos cambiar el punto de corte.

- Sensibilidad (% de bien clasificados en el grupo 1) = 74.3%
- Especificidad (% de bien clasificados en el grupo 0) = 95.6%

Se busca el nivel de probabilidad (p) que maximiza a la vez la sensibilidad y la especificidad, dicho valor se halla en p=0.257.

Tabla de clasificación					
Nivel de prob	**Porcentajes**				
	Correcto	**Sensibilidad**	**Especificidad**	**Falso Positivo**	**Falso Negativo**
0.257	88.4	88.3	88.4	29	4.1

8.6 Tabla de clasificación en el punto optimo

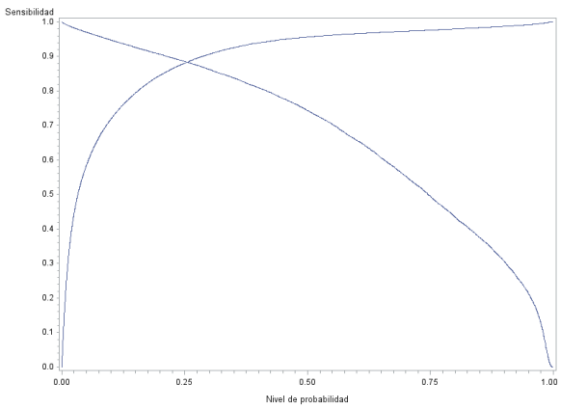
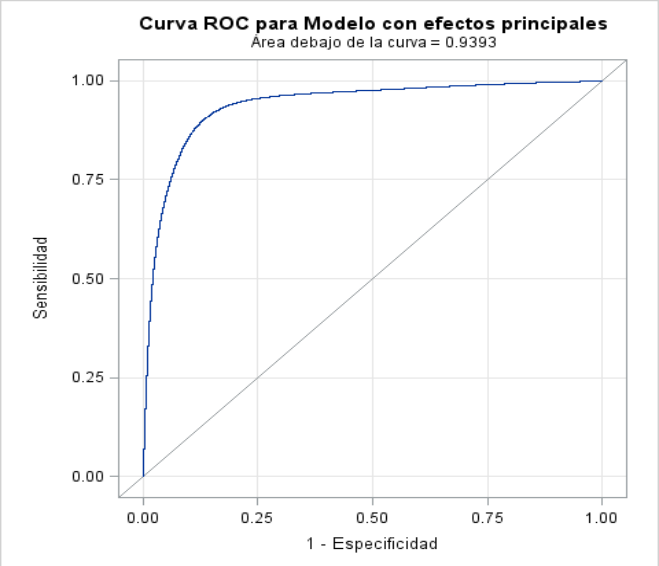


Figura 8.5 Punto de corte óptimo

Representamos el gráfico de sensibilidad y especificidad para todos los valores de p, el punto de coincidencia, será el punto de corte óptimo.

- **Área bajo la curva ROC**



La curva ROC representa la relación entre la Sensibilidad (probabilidad de que el modelo clasifique correctamente el valor de la variable analizada asociada a los individuos que no están emancipados) y 1-Especificidad (probabilidad de que el modelo no clasifique correctamente el valor de la variable analizada asociada a la emancipación) obtenidas para distintos puntos de corte.

Figura 8.6 Área bajo la curva ROC

La calificación del modelo, se presenta mediante el estadístico 'C', es un valor real situado entre 0 y 1. Este será más exacto cuanto más próximos se encuentre de 1. En el modelo el estadístico 'C' toma un valor 0.939, por consiguiente se considera muy buen modelo en términos de poder predictivo.

Asociación de probabilidades predichas y respuestas observadas			
Concordancia de porcentaje	93.9	**D de Somers**	0.879
Discordancia de porcentaje	6	**Gamma**	0.88
Porcentaje ligado	0.1	**Tau-a**	0.323
Pares	20319957450	**c**	0.939

Tabla 8.7 Estadísticos de ajuste

Una vez evaluado el modelo, se quiere comparar los resultados de clasificación que se han obtenidos balanceando la muestra.

Se observa que basar el punto de corte en las probabilidades a priori altera las probabilidades predichas a posteriori de pertenecer a la clase de interés. Ya que una vez equilibrados la sensibilidad y la especificidad, la proporción de falsos positivos aumenta aproximadamente el doble (Tabla 8.7). En cambio, si probamos ajustar el modelo con una muestra equilibrada (50%-50%), se obtienen los resultados de la siguiente tabla, en la cual se tiene aproximadamente el mismo porcentaje de falsos positivos y negativos.

Tabla de clasificación					
Porcentajes	**Porcentajes**				
Nivel de prob	**Correcto**	**Sensibilidad**	**Especificidad**	**Falso Positivo**	**Falso Negativo**
0.5	88.9	85.6	91.2	12.3	10.3

8.7 Tabla clasificación muestra balanceada

Los resultados de la significatividad de los parámetros apenas sufren alteración al ajustar el modelo balanceado. Los objetivos de este apartado se han centrado principalmente en determinar los factores que influyen en la emancipación, así como, definir la forma en que afectan. En el siguiente apartado se hará más hincapié en hallar el modelo con menor tasa de fallos.

En resumen, el modelo pone de manifiesto aspectos que ya habíamos observado con el análisis descriptivo como son la relación positiva entre edad y emancipación o el sexo. De igual forma, se muestra cómo el tener empleo influye de forma positiva en la decisión de emancipación, mientras que vivir en comunidades autónomas con precios elevados de la vivienda dificulta dicha emancipación.

9 Régimen de propiedad entre los jóvenes emancipados

La cultura de la propiedad inmobiliaria en España ha experimentado un descenso de cuatro puntos porcentuales en la última década[4], interrumpiendo así la tendencia al alza registrada desde el censo 1981. Esta preferencia se presenta también en los países del este (Rumania, Croacia, Bulgaria, Lituania, y los países bálticos) que son los únicos que superan a España en cuando al régimen de tenencia en propiedad. En cambio, los países nórdicos son más propensos a la cultura del alquiler como se puede apreciar en la siguiente figura[5].

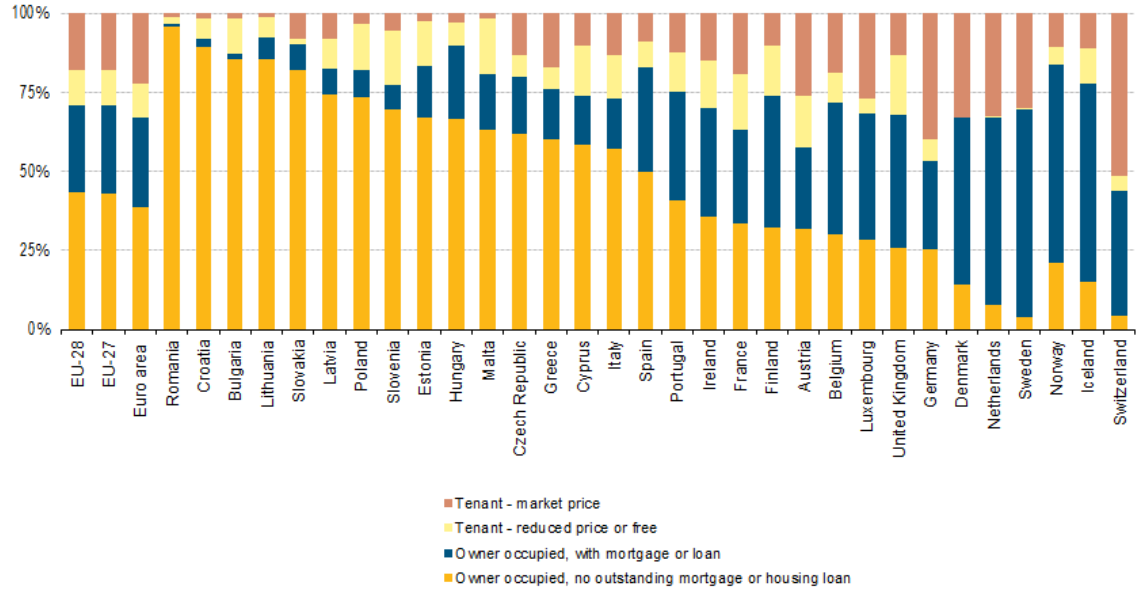


Figura 9.1 Distribución de la población según el régimen de tenencia, 2011.

[4] Proporción de vivienda en propiedad censo 2001: 82.2% frente a 78.9 censo 2011.
[5] Fuente: Eurostat (Oficina de estadística de la UE).

Esta preferencia por la cultura de la propiedad también se aprecia en la población objeto de estudio. Este hecho podría imputarse de algún modo al elevado precio del alquiler, al grifo abierto de crédito hipotecario durante los años de la burbuja inmobiliaria.

En la tabla 9.1 se puede observar que el régimen de propiedad de la vivienda en la que habitan los jóvenes independientemente, casi en la mitad de los casos (49.85%) corresponde a una vivienda en propiedad por compra, con pagos pendientes (hipoteca). Un tercio de jóvenes (29.31%) declara disponer de la vivienda en alquiler. Y un (5.3%) reside en una vivienda propia totalmente pagada y un 3.66% lo hace en una vivienda en propiedad por herencia o donación.

Regimen de tenencia	Frecuencia	Porcentaje
Propia pagada	6084	5.3
Propia (hipotecas)	57209	49.85
Propia por herencia o donación	4204	3.66
Alquilada	33639	29.31
Cedida gratis o a bajo precio	5783	5.04
Otra forma	7844	6.83

Tabla 9.1 Régimen tenencia jóvenes emancipados

Este epígrafe tiene la finalidad de pronosticar el régimen de tenencia de los jóvenes emancipados mediante las técnicas de aprendizaje automático, con el objetivo de comparar los resultados de cada técnica y averiguar cuáles de ellas se ajusta mejor a nuestros datos al cometer menor tasa de fallos. Para llevar a cabo nuestro propósito se tendrá en cuenta únicamente la categoría de tenencia en propiedad con pagos pendientes y el de alquiler, con un 62.97% y un 37.03% respectivamente, siendo la muestra total de 90.848 registros.

Regimen de tenencia	Frecuencia	Porcentaje
Propia (hipoteca)	57209	62.97
Alquilada	33639	37.03

Tabla 9.2 Categorías empleadas en la clasificación

No se ha tenido en cuenta las cuatro restantes categorías ya que apenas alcanzan el 20% entre las cuatro y tampoco son representativas como las categorías consideradas. Cabe mencionar que no se ha trabajado con la muestra entera de 90.848, ya que como se comentó en el apartado de la fase de muestreo 5.1 de SEMMA, algunas técnicas resultan lentas a la hora de la ejecución dado el volumen de los datos. Por este motivo, nos hemos visto obligados a reducir el tamaño muestral para poder obtener resultados de este bloque. En los anexos se presentaran los descriptivos de la muestra extraída. (Tabla A6.1, Anexo 6)

9.1 Clasificación según Régimen Tenencia - Técnicas de Aprendizaje Automático

El aprendizaje automático es una de las áreas de la inteligencia artificial. Es el estudio y desarrollo de modelos cuantitativos que permite a un ordenador realizar distintas tareas sin estar explícitamente programado para hacerlo. El aprendizaje en este contexto equivale a reconocer formas complejas y tomar decisiones inteligentes, de forma que, les permitan resolver nuevos problemas o mejorar su comportamiento en problemas ya tratados.

Antes de desarrollar las técnicas de aprendizaje automático que se van a emplear para la clasificación según el régimen de tenencia, vamos ahondar en unos cuantos aspectos que afectan a todos los métodos que se van a aplicar.

9.1.1 Muestra desequilibrada

Prácticamente cualquier clasificador es sensible a las muestras desequilibras. Estas se caracterizan por tener un número de observaciones menor en las clases de interés, este hecho hace que se dificulte su identificación, o bien presentan un alto número de patrones ruidosos. Estas circunstancias dificultan la construcción de clasificadores, que pueden bien centrarse exclusivamente en la clase mayoritaria o bien no ser capaces de evitar la influencia de los patrones ruidosos.

El caso que se está abordando, no presenta mucho desequilibrio, ya que la clase de interés tiene el 37% de jóvenes. Pero se tendrá en cuenta algunas de las siguientes consideraciones en la construcción del modelo.

- En primer lugar nos podemos plantear el punto de corte según las proporciones muestrales, basándonos en alguna medida de interés como puede ser el caso del estadístico ‘AUC’ Área bajo la curva ROC.
- En la construcción del archivo de entrenamiento, conseguir un equilibrio mayor de las clases (oversampling).

9.1.2 Validación cruzada

Para determinar la validez de los modelos que iremos ajustando con los distintos procedimientos haremos uso de la técnica de validación cruzada [14] que nos resultará útil a la hora de desarrollar y ajustar los modelos, con la finalidad de garantizar la robustez de cada modelo. Para llevar a cabo la técnica, habría que dividir los datos en 3

conjuntos, atendiendo a que cada uno de ellos mantenga la representatividad de la población origen, en la Figura 9.1 se muestra la estructura de la técnica:

- **Conjunto de entrenamiento** (training set): Este conjunto se usa para ajustar los parámetros durante la fase de entrenamiento.

- **Conjunto de verificación** (test set): Con estos datos decidiremos cuando parar el proceso de entrenamiento. Como criterio general, el entrenamiento debe pararse cuando el error del conjunto de verificación sea mínimo.

- **Conjunto de validación** (validation set): Mediante este set se obtienen las correspondientes predicciones con datos que no se han utilizado en el entrenamiento ni en la validación cruzada, evitando así el sobreajuste.

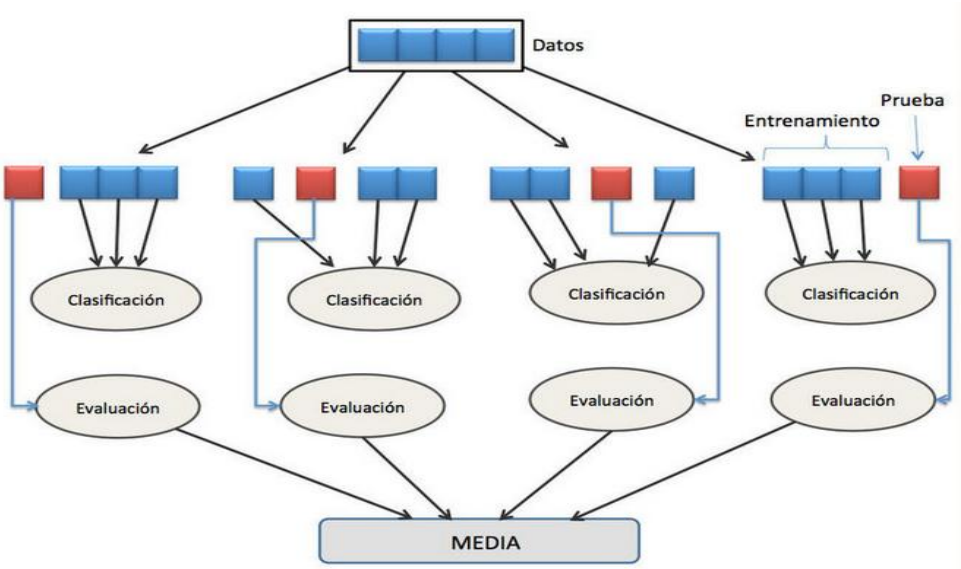


Figura 9.2 Estructura validación cruzada

9.1.3 Selección de variables

En muchas circunstancias se posee información proveniente de una gran cantidad de variables y es preferible seleccionarlas de tal forma, que se elija el subconjunto de ellas que conjuntamente posea más poder discriminante y que no sean redundantes. Para llevar a cabo esta selección, la decisión se tomara apoyándonos en los criterios estadísticos. En la tabla 9.3 se muestran las variables que se considerarán para el modelo de clasificación. Por un lado, se encuentran las características propias de la persona, y por otro lado las relacionadas con la vivienda.

Variables de modelo de clasificación		
Sexo	Binaria	
Nacionalidad	Binaria	Española/ Extranjera
Edad	Intervalo	
Nivel de estudios	Nominal	Primarios/ Secundarios /Bachiller / FP / Universitarios
Estado civil	Nominal	Soltero/ Casado / Separado
Actividad	Nominal	Ocupado / Paro / Otra situación
Comunidad Autónoma	Nominal	6 grupos de clusters obtenidos en el apartado anetrior
Estructura hogar	Nominal	Hogar con una persona sola / Monoparental / Pareja sin hijos / Pareja con hijos / Otra situación
Hijo	Binaria	Sin hijos / Tener uno o más
Ratio_Ocupados		Nº ocupados del hogra / Nº miembros
Aréa	Binaria	> 20.000 Hab / < 20.000 Hab
Superficie	Intervalo	
Nº Habitaciones	Intervalo	
Estado vivienda	Binaria	Malo / Bueno
Año construcción vivienda	Intervalo	
IPV	Intervalo	Índice precio de la vivienda

Tabla 9.3 variables consideradas para la clasificación

A continuación se muestran dos gráficos en los cuales se ordena las variables explicativas según su importancia en cuanto a la variable dependiente. Los dos criterios que se han empleado para la selección de variables son el criterio Chi cuadrado y el índice de Gini.

Criterio Chi cuadrado: Se construye la tabla de contingencias en este caso de la variable a predecir tipo de tenencia (Alquiler/ Hipoteca) frente a cada una de las variables regresoras categóricas, con el fin de calcular el valor del estadístico chi-cuadrado. Si la significatividad vinculada al estadístico es menor, aumenta la importancia de la variable regresora respecto a la variable objetivo.

Se procede a ordenar las variables regresoras con mayor valor del estadístico asociado, cruzando la variable dependiente con la independiente. Vemos en la Figura 9.3 que la nacionalidad obtiene el mayor valor del estadístico. Por lo tanto podemos decir que la nacionalidad tiene mayor relevancia con respecto al tipo de tenencia. Las siguientes variables que se encuentran en este ranking son estructura del hogar y la comunidad autónoma.

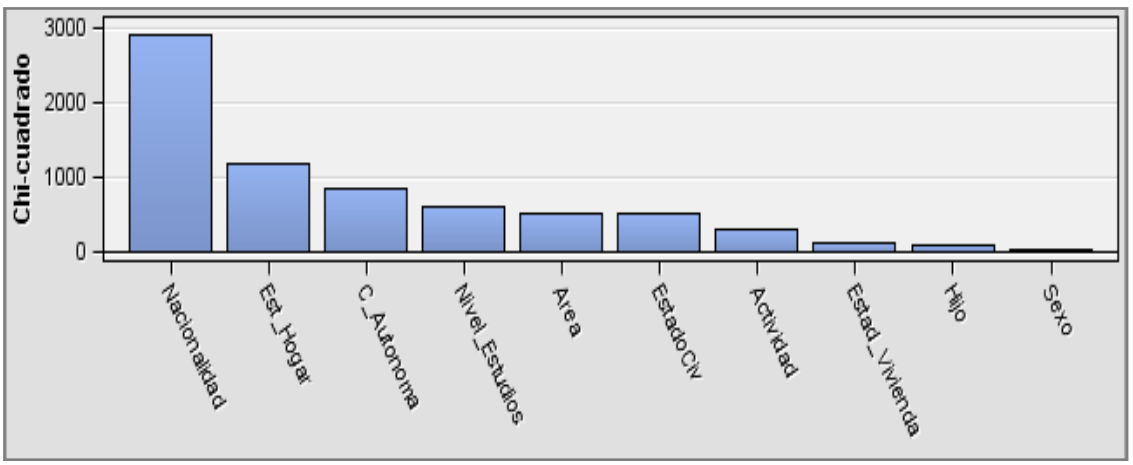


Figura 9.3 Orden de importancia por el criterio Chi cuadrado

Índice de Gini: El estadístico de Gini es uno de los estadísticos más utilizados para evaluar las diferencias entre dos poblaciones. Para el caso que nos atañe, se medirá las diferencias en la población de 'Alquiler' y 'Hipoteca'.

En la figura 9.4 se ordena las variables en función del valor del estadístico Gini que se muestra en tantos por cien. Vemos que la nacionalidad vuelve a ocupar el primer puesto, confirmando así los resultados obtenidos con el criterio anterior. Adicionalmente se observa como las variables de intervalo años de construcción y superficie de la vivienda se encuentran en las primeras posiciones.

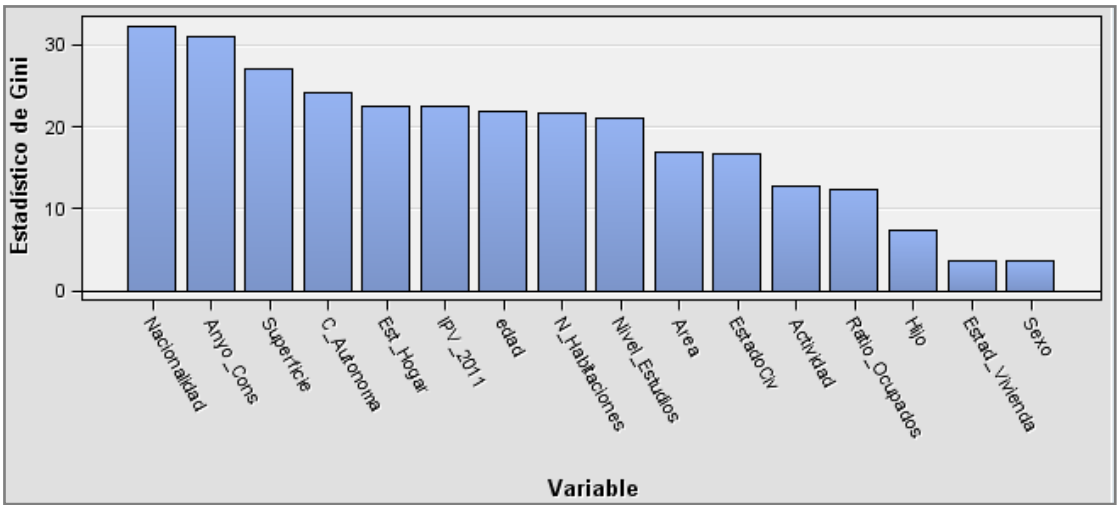


Figura 9.4 Orden de importancia por el criterio de Gini

Una vez definidos los criterios que se han empleado para tener un primer contacto con las variables que muestran tener mayor poder discriminante, se puede decir que las variables nacionalidad, años de construcción, superficie, estructura hogar y comunidad autónoma tienen mayor poder predictivo que el resto. Pero esta coyuntura no nos llevará a emplear solamente dichas variables, sino que se irá probando con varias combinaciones que se irá seleccionando mediante modelos hasta que se obtenga el mejor modelo.

En los siguientes bloques se desarrollarán las técnicas de aprendizaje automático que se han llevado a cabo, el resultado con el mejor modelo de cada técnica se expondrá en el apartado 9.5 de comparación de técnicas. Para la elección de los modelos, así como para comparar los distintos métodos, se tendrá en cuenta el estadístico (AUC) el área bajo la curva ROC y la tasa de fallos que viene dada por:

Tasa fallos= 1- $\left(\frac{Verdaderos\ positivos+Verdaderos\ negativos}{n}\right)$, siendo n el tamaño muestral.

9.2 Regresión logística binaria

Se vuelve a emplear el modelo de regresión logística para llevar a cabo el objetivo que se tiene de clasificar a la muestra de jóvenes emancipados según el régimen de tenencia de la vivienda. Dado el propósito que se quiere alcanzar, se van a presentar de forma breve los odds ratio de las variables regresoras que resultaron ser significativas en el modelo. De esta manera podremos cuantificar la influencia de cada variable regresora, este hecho hace que la regresión logística este en ventaja frente a las redes neuronales, ya que los pesos estimados de esta última no son interpretables, y en cuanto a los árboles de clasificación se puede obtener la importancia de cada variable.

Todas la variables que se han considerado para ajustar los modelos de este bloque (tabla 9.3) resultaron ser significativas a excepción de la variable hijos (tener o no hijos) y el estado de la vivienda (bueno/malo), coincidiendo así con los resultados de los dos criterios de selección de variables.

En la tabla 9.4 se presentan los Odds Ratio asociados al modelo logístico, siendo la variable objetivo régimen de tenencia de la vivienda (Hipoteca/ Alquiler), de forma resumida se explican los Odds Ratio:

Efecto	Categoria	Estimador del punto	Límites de confianza al 95% de Wald	
Nivel Estudios (Ref=Primarios)	Secundarios	0.85	0.746	0.969
	Bachiller	1.251	1.075	1.455
	FP	0.845	0.732	0.975
	Universitarios	1.627	1.407	1.881
Sexo (Ref=Hombre)	Mujer	0.804	0.744	0.869
Area (Ref= > 20.000 Hab)	< 20.000 Hab	1.252	1.155	1.357
Estructura del hogar (Ref= Hogar con una persona sola)	Monoparental	0.689	0.517	0.917
	Pareja sin hijos	0.634	0.554	0.725
	Pareja con hijos	0.581	0.476	0.709
	Otra situación	1.172	0.975	1.409
Actividad (Ref=Ocupados)	Parado	1.042	0.921	1.18
	Otra situación	0.855	0.724	1.009
IPV_2011		0.844	0.782	0.911
EstadoCiv (Ref= Soltero)	Casado	0.484	0.443	0.529
	Separado/Divorciado	0.79	0.613	1.017
Nacionalidad (Ref=Española)	Extranjera	6.96	6.26	7.737
C_Autonoma (Ref= País Vasco, Navarra)	Andalucía, Extremadura, Murcia, C.Valenciana y C. la Mancha	0.728	0.603	0.88
	Cantabria, Cataluña y la Rioja	2.558	2.047	3.196
	C. León, Asturias, y Galicia	1.347	1.067	1.701
	Aragón y Madrid	1.893	1.574	2.276
	Canarias y Baleares	1.221	0.935	1.595
N_Habitaciones		0.819	0.791	0.847
Ratio_Ocupados		0.563	0.477	0.664
Superficie		0.993	0.991	0.994
Año construcción		0.835	0.818	0.851
Edad		0.857	0.842	0.872

Tabla 9.4 Odds Ratio estimados

Se observa que los siguientes aspectos aumentan el riesgo de estar en una vivienda en régimen de alquiler: Tener nacionalidad extranjera, vivir en municipios con menos de 20.000 habitantes y tener estudios universitarios, entre otras. En cambio, el hecho de vivir en pareja con o sin hijos, disminuye el riesgo de estar en una vivienda en régimen de alquiler con respecto a una persona que vive sola, así como el hecho de estar casado.

En cuanto a las variables regresoras relacionadas con la vivienda, se observa que a medida que se reduce el índice de precio de la vivienda, disminuye el riesgo de estar en una vivienda en régimen de alquiler. Y en cuando al número de habitaciones, superficie, y año de construcción se aprecia que aumentando estos aspectos disminuye el riesgo de encontrarse en régimen de alquiler.

Por último, se obtiene que los jóvenes emancipados de las comunidades autónomas de Cantabria, Cataluña y la Rioja tienen mayor riesgo de estar en régimen de alquiler frente a los jóvenes vascos y navarros.

En el apartado 9.5 de comparativa de técnicas, se describirá el modelo de regresión logística con el cual se ha obtenido la menor tasa de error al clasificar. (En el anexo se presentan los resultados con la estimación de los parámetros, Tabla A6.2, Anexo 6).

9.3 Diseño y entrenamiento de modelo de red neuronal

El entrenamiento de una red neuronal, tiene por objetivo modificar los pesos de la red con la finalidad de que coincida la salida deseado por el usuario con los resultados obtenidos por la red mediante un determinado criterio de entrada (Apartado 3.4). En otras palabras, el entrenamiento de una red neuronal es un problema de minimización no lineal en el cual los pesos de la red son iterativamente modificados con el fin de minimizar el error entre la predicción y la respuesta esperada.

La red neural busca minimizar una función de error la cual se evalúa en los nodos de salida. En aras de minimizar dicha función, se pueden aplicar diversos algoritmos de optimización.

El entrenamiento de la red se hará en tres fases (elección nodos, algoritmo de optimización, y función de activación), este orden no está determinado, se ha estructurado de esta manera para mantener un cierto orden.

9.3.1 Determinación del número de nodos ocultos

Se empieza la fase de aprendizaje o entrenamiento de la red neuronal con todo el conjunto de variables de la base de datos, en dicha fase la red es entrenada mediante algoritmos con el objetivo de realizar clasificaciones sobre nuevos conjuntos test.

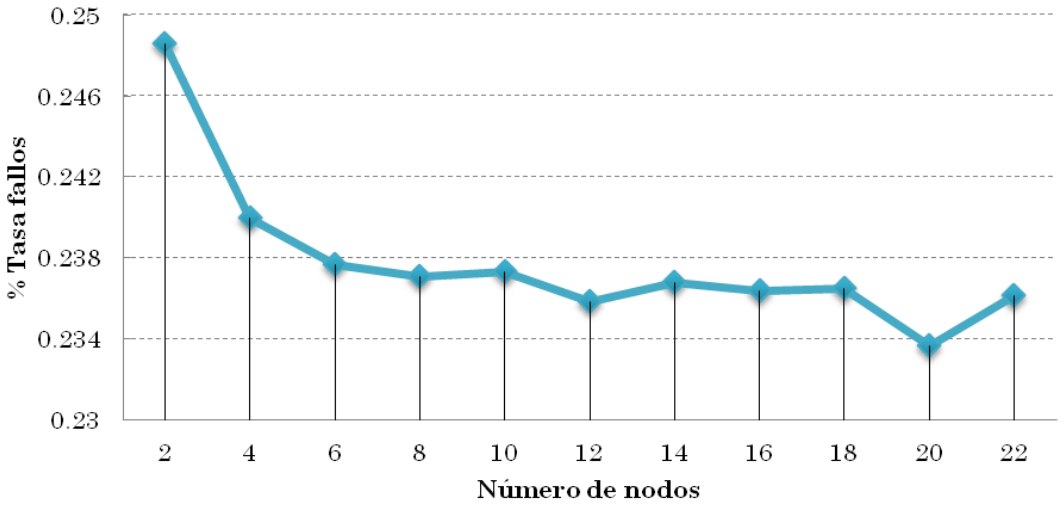


Figura 9.5 Evolución tasa fallos según número nodos

Para obtener una visión global sobre el número de nodos que tenemos que emplear, se plasma en la figura 9.5 la evolución de número de nodos respecto a la tasa de fallos (El algoritmo de entrenamiento y la función de activación de momento los fijaremos por la opción por defecto del SAS, esto es 'Levenberg–Marquardt' y 'Softmax' respectivamente.). A simple vista se puede apreciar la tendencia decreciente en la tasa de fallos a medida que se aumenta el número de nodos, manteniéndose casi constante a partir de los 10 nodos. Este hecho no hay que tomarlo al pie de la letra, ya que el resultado esta obtenido con una sola partición y sin validación cruzada.

Se obtienen resultados muy parecidos mediante la validación cruzada y con distintas semillas aleatorias que inicia el proceso, ya que esto puede alterar en gran medida los resultados. Esto reafirma de algún modo el resultado obtenido en la figura 9.5 sin variar las semillas. Después de realizar la operación un par de veces, llegamos a la conclusión de que el intervalo óptimo de nodos que aplicaremos está entre 6 y 10 nodos.

Una vez determinado el intervalo de nodos sobre el cual vamos a entrenar la red, pasamos a determinar el algoritmo de aprendizaje, este último es un proceso por el cual la red neuronal modifica sus pesos en respuesta a una información de entrada.

9.3.2 Algoritmos para la optimización

Los algoritmos propuestos para el aprendizaje de redes se caracterizan por ser de aprendizaje supervisado, lo cual permite el cálculo de los parámetros de la función que aproxima la salida deseada, que se mide a través de una función de error de ajuste. En la siguiente figura se ajustan cinco modelos de redes neuronales cada uno de ellos con un algoritmo distinto, el número de nodos se ha ido variando de 6 a 10 nodos, y de momento se sigue trabajando con la función de activación la 'Softmax'.

A continuación se citarán los algoritmos de optimización que se han empleado para el entrenamiento de la red, junto con la abreviatura que se utilizará en los gráficos y la correspondiente referencia para consultar los fundamentos de cada método.

- Método Levenberg–Marquardt 'LEVMAR' [18].
- Algoritmo Backpropagation 'BPROP' [25]
- Algoritmo Quasi-Newton 'QUAEW' [25]
- Algoritmo Gradiente Conjugado Conjugate Gradients 'CONGRA' [17]
- Método Región de confianza Trust Region 'TRUREG' [8]

En la figura 9.6 se evalúan los resultados de cada algoritmo, mediante el área bajo la curva (estadístico AUC) para cada fichero.

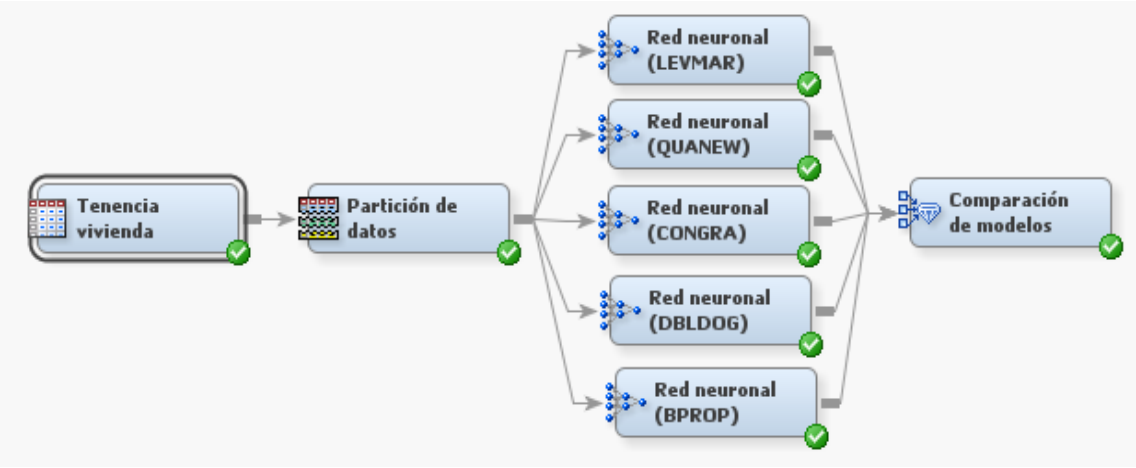
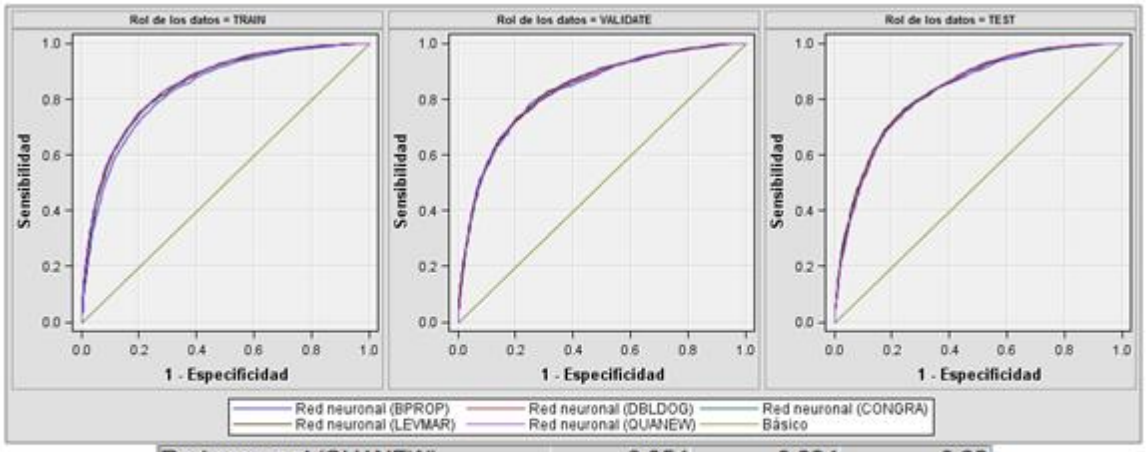


Figura 9.6 Selección algoritmo de optimización

En la siguiente figura, se muestra el resulta de los cinco modelos ajustados. Como es de esperar el estadístico C presenta valores bajos en el fichero de prueba, ya que son datos que no han participado en el ajuste del modelo. A tenor de lo que observamos se podría decir que el mejor resultado lo obtiene el algoritmo de optimización 'Levenberg–

Marquardt, pero al igual que hicimos en el anterior apartado habrá que corroborar estos resultados mediante la validación cruzada.



Red neuronal (QUANEW)	0.854	0.834	0.83
Red neuronal (LEVMAR)	0.851	0.834	0.833
Red neuronal (DBLDOG)	0.851	0.833	0.831
Red neuronal (CONGRA)	0.849	0.834	0.832
Red neuronal (BPROP)	0.836	0.83	0.828

Figura 9.7 Área bajo la curva ROC

En la figura 9.6, se muestra la distribución de la tasa de fallos media por algoritmo, todos ellos presentan resultados similares. En cambio, el algoritmo de ***Back propagation*** basado en la propagación del error hacia atrás en la red, presenta un error bastante alto respecto al resto de algoritmos. Este aspecto podría ser debido a los parámetros *learning rate* y *momentum* que ponderan la variación en los pesos en cada iteración de tal forma que el algoritmo aprenda del error o de la variación de estos en anteriores etapas respectivamente. Por lo que nos planteamos variar estos parámetros y tenerlos en cuenta en los resultados finales.

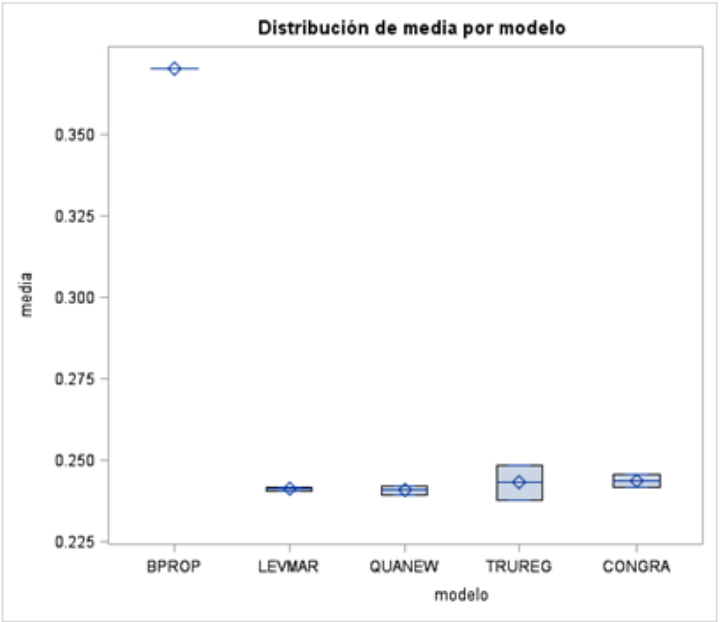


Figura 9.8 Resultado validación cruzada

9.3.3 Función de activación

Una vez definido el algoritmo de optimización para la estimación de los pesos, habría que determinar la función de activación que fija el nivel de activación de salida de la neurona.

Dentro de los parámetros que definen una red, la función de red más utilizada es de tipo lineal, y como función de activación más empleada está la función sigmoidea. Estas son las funciones que se han probado para nuestros datos:

$$y = f\left(w_{j,out}H_j + \ \ldots + w_{1,out}H_1\right) + \ b_{out} = f\left(\sum_{j}^{m} w_{j,out}H_j \ + \ b_{out}\right) = f(a)$$

- Función Sigmoide $f(a) = \frac{1}{1+e^{-a}},\ 0 < a < 1$
- Función Tangente hiperbólica $f(a) = \ 1 - \frac{2}{(1+e^{(2a)})},\ -1 < a < 1$
- Función Arcotangente $f(a) = \arctan(a) * \frac{2}{\pi},\ -1 < a < 1$
- Función Gausiana $f(a) = e^{-\frac{(a-c)^2}{2b^2}}, 0 < a < 1,\ b > 0$
- Función SINe $f(a) = \sin(a), 0 < a < 1$
- Función Softmax $f(a) = \frac{e^a}{\sum_{i=1}^{m} e^{x^T w_m}},\ 0 < a < 1$

Se van a mostrar los resultados de la tasa media de fallos mediante validación cruzada repetida (distintas semillas), combinando para ello, los algoritmos de optimización y las funciones de activación.

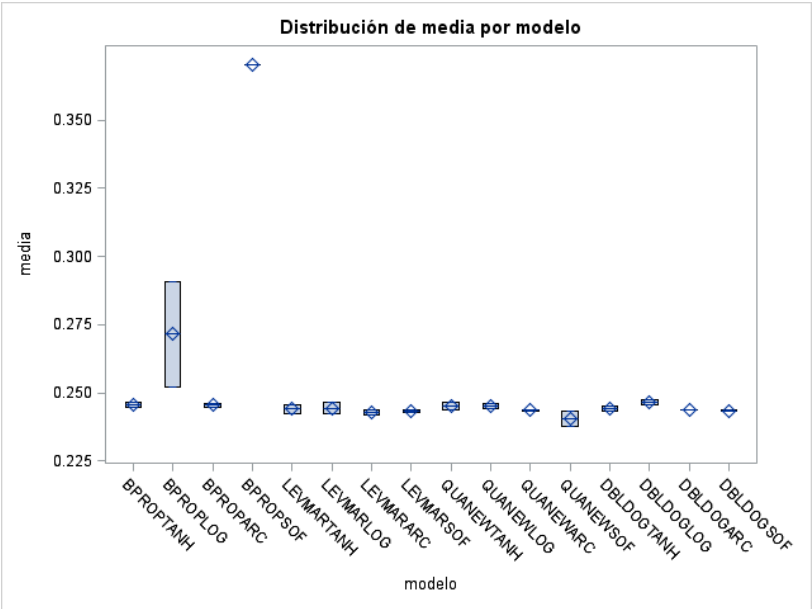


Figura 9.9 Resultados validaos cruzada algoritmo con función de activación

De los resultados de la figura 9.8, se desprende que las únicas alteraciones que se observan corresponden al algoritmo '***Back propagation'*** cuando se aplica la función de activación '***Softmax'***.

Un vez explicado todo el proceso para entrenar un modelo de red neural, se ha procedido a controlar los parámetros de la red para distintos set de variables, con el fin de obtener el menor error, dado por la tasa de fallos mediante la validación cruzada repetida, En el apartado de comparaciones se presentaran los resultados finales.

9.4 Árboles de clasificación

En los siguientes bloques se desarrollan los métodos Bagging y Random Forest basados en los modelos de árboles de clasificación (Apartado 3.5), previamente se describirá el método de remuetreo Bootstrapping que se empleará en las dos técnicas.

9.4.1 Técnica Bootstrapping

El Bootstrapping es un método de remuetreo (Efron, 1979) [12] que consiste en generar un elevado número de muestras con el objetivo de estudiar el comportamiento de un determinado estadístico. Las muestras extraídas tienen las siguientes características

- Muestras con reemplazamiento, los individuos pueden repetirse en el mismo set de datos, teniendo cada registro una probabilidad de $\frac{1}{n}$ de ser escogido cada vez.
- El tamaño de la muestra es igual al tamaño de los datos de entrenamiento, pero la diferencia radica en que no son los mismos registros, al ser un muestreo con reemplazamiento, ya que habrá registros repetidos.
- Muestra uniforme, mismos registros en cada muestra.

Una vez extraído un número elevado de muestras, se calculará el valor del estadístico de interés en cada muestra (en nuestro caso será la tasa de fallos), que se empleará como estimador del parámetro poblacional. Se determina la función de distribución empírica del parámetro de interés, que representa una buena aproximación a la verdadera distribución de probabilidad del estadístico.

9.4.2 Algoritmo Bagging

El algoritmo Bagging (Breiman, 1996) [5] se construye combinando la salida de varios clasificadores con el objetivo de generar uno más potente. Emplearemos el algoritmo en

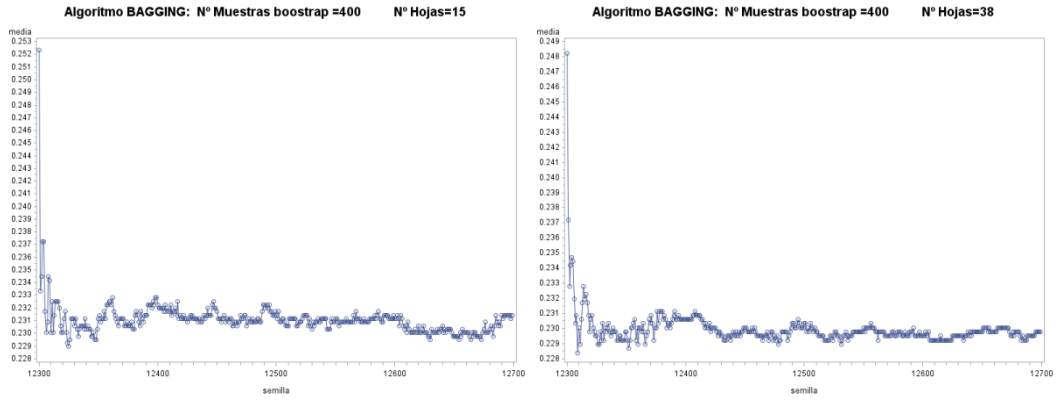
el modelo de árboles de decisión, pero se podría utilizar en otros modelos como puedes las redes neuronales.

La idea de Bagging es generar a partir de un conjunto de entrenamiento de tamaño N, *m* nuevos conjuntos de entrenamiento. Estos nuevos conjuntos se construyen con la generación de muestras Bootstrap. La técnica de Bagging sigue estos pasos:

- Seleccionar N observaciones de los datos originales. Obteniendo como resultado diferentes muestras Bootstrap.
- Se crea un modelo predictivo con cada muestra, obteniendo *m* modelos diferentes.
- Se construye un único modelo predictivo, que es el promedio de los *m* modelos.

Si el modelo predice una salida numérica, el algoritmo Bagging se crea promediando las salidas de los distintos clasificadores. Si el modelo es de clasificación como resulta ser nuestro caso, la salida del clasificador combinado será aquella clase que resulte ser elegida por la mayoría de los *m* clasificadores.

A continuación se llevara a cabo el procedimiento que se ha descrito. El objetivo sigue siendo el parámetro 'media' (tasa de error) que se presenta en el siguiente gráfico, siendo esta la función que se quiere minimizar. Para ello, se va a probar con varios árboles, donde se irá variando tanto el número de hojas como el tamaño de esta.



Variable de análisis : Error (Fichero validate)				
N	Media	Dev std	Mínimo	Máximo
3634	0.2325261	0.4225007	0	1
Variable de análisis : Error (Fichero Train)				
N	Media	Dev std	Mínimo	Máximo
14536	0.2346588	0.4237999	0	1

Variable de análisis : Error (Fichero validate)				
N	Media	Dev std	Mínimo	Máximo
3634	0.2300495	0.4209222	0	1
Variable de análisis : Error (Fichero Train)				
N	Media	Dev std	Mínimo	Máximo
14536	0.2291552	0.4203038	0	1

Figura 9.10 Resultados algoritmo Bagging

En la figura 9.10 se muestran dos ejemplos, en los cuales se ha extraído 400 muestras Bootstrap, la diferencia entre los dos ejemplos radica en el número de hojas de cada árbol, se aprecia que el árbol con mayor número de hojas (árbol agresivo), muestra una

tasa de fallos menor que el árbol con menos hojas (árbol débil). Una vez visto el procedimiento, se tiene que los parámetros que se tienen que monitorizar son, el número de hojas del árbol, y el número de muestras Bootstrap, esta última se podría fijar en aproximadamente 100 muestras, ya que se ve que el proceso de minimización converge, como se puede apreciar en la figura anterior.

Los resultados obtenidos en la anterior figura, no hay que tomarlos al pie de la letra, ya que solamente están basados en una partición (entrenamiento y validación), para afianzar estos resultados se va a emplear validación cruzada repetida, para ello se va a probar varios set de variables al igual que se hizo en el apartado de las redes neuronales.

9.4.3 Algoritmo Random Forest

Random forest (Breiman, 2001) [6] es una técnica mejorada de Bagging, este algoritmo mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en las variables utilizadas para segmentar cada nodo del árbol. El proceso sigue los mismos pasos que Bagging, la única diferencia radica en que, en cada modelo predictivo que se ajusta, se seleccionarán en cada nodo *p* variables de las *k* originales, y de las *p* elegidas, se escogerá la mejor variable para la partición del nodo, obteniendo así en cada modelo, diferentes registros y diferentes variables.

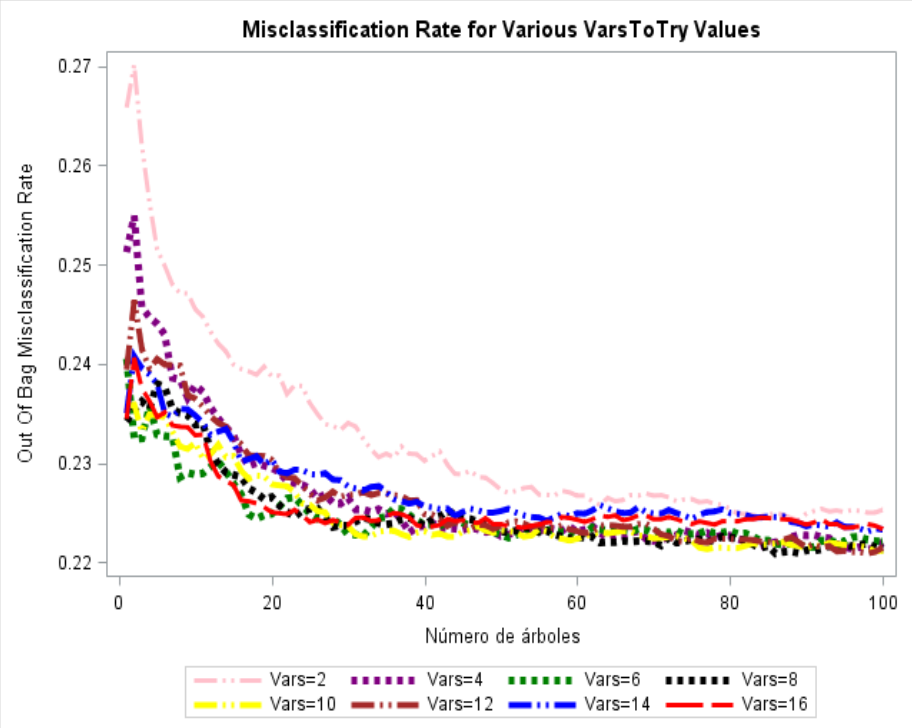


Figura 9.11 Tasa fallos en función de número de muestras

Una vez explicada la metodología que sigue el algoritmo Random Forest, se presenta en la figura 9.11 la tasa de fallos en los datos de validación en función de número de árboles, variando el número de variables a tener en cuanta en cada nodo a la hora de hacer la partición.

Se puede apreciar que el hecho de seleccionar solamente dos variable en cada set de datos, presenta mayor tasa de fallos. En cuanto se pasa de 2 a 4 variables disminuye considerablemente el error medio en los primeros 50 árboles. En cuanto al número de árboles, vemos que a partir de 40 muestras el porcentaje de error se mantiene aproximadamente constante para las distintas combinaciones de variables. Se ha repetido el proceso varias veces para determinar el número de variables óptimo, así como el número de árboles. Como se ha detallado en el proceso, los parámetros que hay que monitorizar del algoritmo Random Forest, son los que se han explicado en el método Bagging, más el número de variables a entrenar.

En cuanto a la importancia de las variables, se ha visto que la nacionalidad, Estado civil y superficie son las que más aportan a las divisiones realizadas. Coincidiendo en cierta forma con los procesos de selección de variables presentados mediante los dos criterios Índice de Gini y el estadístico Chi Cuadrado.

9.5 Comparación de las técnicas de clasificación

Una vez detalladas las técnicas empleadas se procede a comprarlas y a detallar los distintos modelos que se han seleccionado. Mediante este apartado se ha pretendido estudiar la arquitectura que tiene un modelo de red neuronal, así como los dos métodos de combinación de clasificadores Bagging y Random Forest, poniendo de manifiesto tanto a nivel teórico como a nivel experimental la especificación de cada una de las técnicas, con la finalidad de aclarar los parámetros que se tienen que controlar con cada uno de los métodos.

Antes de detallar los modelos, cabe destacar que el punto de corte que se ha ido ajustando para los distintos modelos, se determinó a priori en las proporciones muestrales. A medida que hemos ido controlando los parámetros se ha ido basando el corte en la tasa de fallos apoyándonos en el estadístico AUC de la curva ROC.

A continuación se detalla la estructura del modelo final de cada técnica, el cual obtuvo la menor tasa de fallos media en los distintos ficheros de prueba a la hora de

pronosticar, A la vista de los resultados de la figura 9.12 se puede advertir que el modelo de regresión logística es el que presenta la menor tasa de fallos media en la validación cruzada.

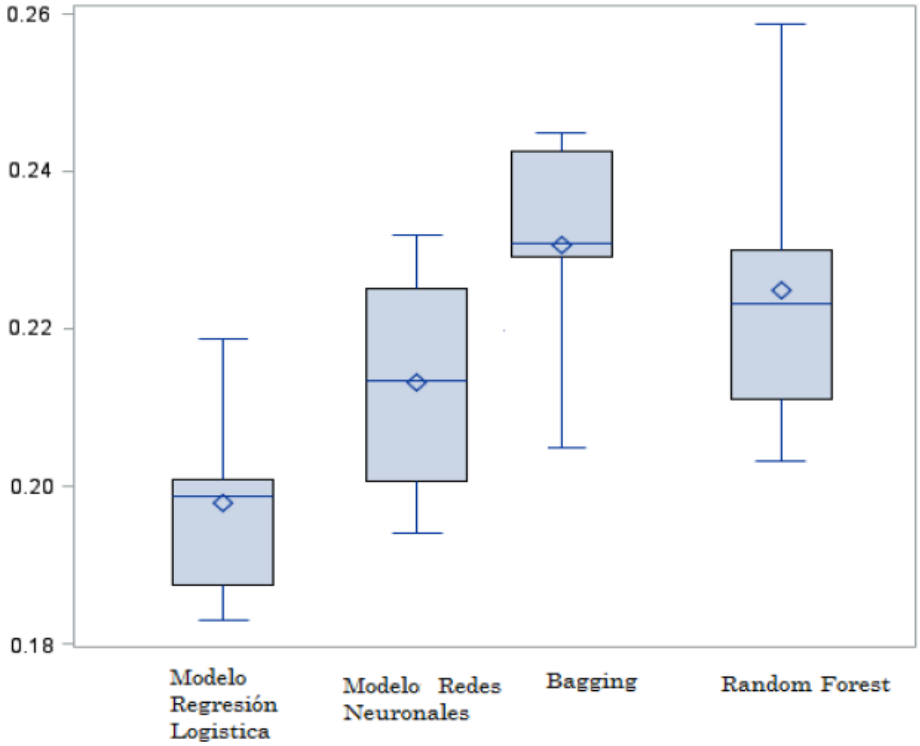


Figura 9.12 Distribución tasa media de fallos por modelo

- Regresión logística: modelo con efectos principales e interacciones, todas las variables regresoras a excepción de las dos variables 'Estado de la vivienda' y 'Tener o no hijos'.
- Redes Neuronales: modelo con 10 nodos, algoritmo 'Levenberg–Marquardt' junto con la función 'Softmax'. Se han probado varios sets de variables, pero finalmente, se optó por las variables que se han utilizado en la regresión logística.
- Método Bagging: Se ha fijado el número de muestras Bootstrap en 35 muestras, y las mismas variables predictoras que las anteriores técnicas.
- Técnica Random Forest: En cuanto a este último método, se ha fijado el número de muestras en 60, y en 10 las variables a entrenar en cada partición, y en cuanto a las variables se han tenido en cuenta todas las variables regresoras.

En la figura se puede apreciar que los dos métodos de agregación Bagging y Random Forest presentan una mayor dispersión en comparación con los otros dos modelos, se realizó el contraste de comparación de medias para determinar si la tasa media de fallos entre las distintas técnicas es similar. De la prueba se concluye que hay diferencias estadísticamente significativas entre los modelos.

10 Conclusiones

En el presente trabajo se han abordado diversos aspectos relacionadas con los jóvenes españoles, abarcando desde el nivel educativo hasta la independencia del hogar familiar. Se ha centrado el foco principalmente en identificar qué factores influyen en mayor o menor medida en las cuestiones analizadas, así como hallar las relaciones subyacentes.

A continuación se sintetizan las conclusiones más destacadas que han arrojado los distintos modelos obtenidos con técnicas de minería de datos. Mediante estos procedimientos hemos podido realizar exploraciones en profundidad, y extraer la información tan valiosa que encierran los datos.

Previamente al ajuste de los modelos, se optó por agrupar las comunidades autónomas en base a datos macroeconómicos mediante un análisis cluster, este análisis arrojó seis conglomerados diferenciados.

En el siguiente epígrafe se estudiaron los factores que inciden sobre la probabilidad de cursar estudios universitarios, destacándose lo siguiente:

- Se ha visto que hay más mujeres con estudios universitarios entre la población de jóvenes de 22 a 30 años de edad. Este hecho se reafirmó en el modelo logístico ordinal, al comprobar que las mujeres tienen mayor probabilidad de contar con estudios superiores.
- Para los jóvenes de padres extranjeros es menos probable que accedan a estudios universitarios.
- El nivel educativo de los padres influye de forma positiva en la educación de los hijos. Los jóvenes de padres con estudios superiores tienen mayor probabilidad de alcanzar estudios universitarios frente a los jóvenes cuyos padres sin estudios superiores.

Una vez estudiados los factores que influyen en el nivel educativo. Se analizó la situación de los jóvenes universitarios ocupados, en cuanto a los cargos que ocupan, empleando para ello la clasificación nacional de ocupaciones. De los resultados obtenidos del análisis estadístico, se señala lo siguiente:

- Los jóvenes con estudios en la rama de ciencias sociales y jurídicas se encuentran vinculados a los cargos de directores y gerentes.

- Los titulados en la rama de artes y humanidades, se hayan relacionados con las ocupaciones elementales y el sector de los servicios de restauración.
- En cuanto al área de ciencias de la salud y la educación, estas especialidades se vinculan con la categoría de técnicos y profesionales científicos e intelectuales.

En el siguiente apartado, se ha pretendido identificar los factores que influyen en el proceso de emancipación, tanto si lo propician como si lo que la retrasan:

- El modelo pone de manifiesto que el hecho de ser mujer, o tener nacionalidad extranjera, o estar casado frente a soltero, favorecen la emancipación.
- Por el contrario, se tiene que los jóvenes con estudios universitarios tienen menor probabilidad de emancipación. Este aspecto se achaca a la tardía terminación de estudios y a la falta de autonomía económica hasta ese momento, frente a los jóvenes con estudios básicos, que se incorporan antes al mercado laboral.

Para finalizar el estudio, se desarrollaron varias técnicas de aprendizaje automático con la finalidad de hallar el mejor modelo predictivo para pronosticar el régimen de tenencia de la vivienda de los jóvenes emancipados. En concreto, se han estudiado las redes neuronales y las dos técnicas de agregación Bagging y Random Forest, junto con regresión logística. De la comparativa de estas técnicas se deprende que el modelo de clasificación con menor tasa de fallos es el de regresión logística. Mediante este modelo se pudo cuantificar la relación entre estar en una vivienda en régimen de alquiler y las variables macroeconómicas, propias del individuo, o características de la vivienda, consideradas en el modelo:

- Los resultados concluyen que ser extranjero, vivir en un municipio de menos de 20.000 habitantes y tener estudios universitarios, aumentan la probabilidad de estar en una vivienda alquilada.
- En cambio, el hecho de vivir en pareja o el estar casado, disminuye la probabilidad de estar en una vivienda de alquiler frente a una persona que vive sola.
- También se ha observado que a medida que los precios de la vivienda son más bajos, se aprecia una mayor propensión a tener la vivienda en propiedad.
- Y respecto a las características de la vivienda, se ha constatado que a mayor tamaño de la vivienda, y a menor antigüedad de su construcción, se tienen mayores niveles de tenencia de la vivienda con hipoteca.

11 Bibliografía

[1] Consejo de la Juventud de España. "*Observatorio de emancipación España*". 4º Trimestre 2014. Disponible en: http://www.cje.org/descargas/cje6176.pdf

[2] Censos de Población y Viviendas 2011. Instituto Nacional de Estadística. Disponible en: http://www.ine.es/censos2011_datos/cen11_datos_inicio.htm

[3] Oficina Europea de Estadística. EUROSTAT. Disponible en: http://ec.europa.eu/eurostat/web/main/home

[4] **Alfaro, E. Gámez, M. García, N.** "*Una revisión de los métodos de agregación de clasificadores*". Área de Estadística. Departamento de Economía y Empresa. Universidad de Castilla-La Mancha.

[5] **Breiman, L.** "*Bagging predictors*". Machine Learning, 1996. 24: p.123-140.

[6] **Breiman, L.** "*Random forests*". Machine Learning, 2001. 45: p. 5–32.

[7] **Brian S. Everitt.** "*Cluster Analysis*". Editorial John Wiley and Sons Ltd. 2009.

[8] **Byrd, B. Schnabel, R. Shultz, G**. "*A Trust Region Algorithm for Nonlinearly Constrained Optimization*". Society for Industrial and Applied Mathematics, 1986. 24: p. 1152-1170.

[9] **Carl J. Huberty**. (2006), "*Applied Manova and Discriminant Analysis*". Editorial John Wiley and Sons Ltd.

[10] **Derr, B.** "*Ordinal Reponse Modeling With the Logistic Procedure*". SAS Institute Inc. p 446. 2013.

[11] **Echaves, A. Andújar, A**. "*Acceso a la Vivienda y Emancipación Residencial de los jóvenes Españoles en un Contexto de Crisis*". XIV Congreso Nacional de Población. 2014.

[12] **Efron, B.** "*Bootstrap methods: Another look at the jackknife*". The Annals of Statistics, 1979. 7: p.1-26.

[13] **Greenacre, Michael**. "*La práctica del análisis de correspondencias*". Rubes Editorial. 2008.

[14] **Gutirrez-Osuna, Ricardo.** "*Leave-one-out Cross Validation*", Wright State University.

[15] **Hastie, T. Tibshirani, T. Friedman, J**. "*The Elements of Statistical Learning*". Springer. Stanford, CA, 2008.

[16] **Jovell, Albert,** "*Análisis de regresión logística*". Editorial Centro de Investigaciones Sociológicas.2006.

[17] **Magnus R. Hestenes and Eduard Stiefel.** "*Methods of Conjugate Gradients for Solving Linear Systems*". Jornal of Research of the National Bureau of Standards.

[18] **Manolis I. A. Loukaris**. "A *Brief Description of the Levenberg-Marquardt Algorithm Implemented by Levmar*". Foundation for Research and Technology - Hellas, 2005.

[19] **Mccullagh, P.** "*Regression models for ordinal data*". Journal of the Royal Statistical Society, 42, 109-142. 1980.

[20] **Michie, D. Spiegelhalter, D.J. Taylor, C.C.** "*Machine Learning, Neural and Statistical Classification*". 1994.

[21] **Pastor, J.M. Peraita, C. Soler, Ángel.** *"Determinantes de la realización de estudios universitarios en España*". XXIV Jornadas de la asociación de Economía de la Educación. 2015.

[22] **Patón, J.M**. "*Emancipación Juvenil y Políticas de Vivienda en Europa*". Arquitectura, Ciudad e Entorno. 2007.

[23] **Refaat, Mamdouh**. "*Data Preparation For Data Mining Using SAS*". Morgan Kaufmann Publishers. 2007.

[24] **Rodríguez, J.E. Barrios, J.A.** "*Vivienda de Protección Oficial Libre: Un Modelo Logit Mixto de Tenencia de Vivienda en Canarias*". Dpto. de Economía Aplicada. Universidad de La Laguna.

[25] **Sarle, W.** "*Neural Network Implementation in SAS Software*". SAS Institute

[26] **Stokes, M.E. Davis, C.S. Koch, G.G**. "*Categorical Data Analysis Using SAS*". Third Edition. 2012.

[27] **Uriel, E**. (2005), "*Análisis Multivariante aplicado*". Editorial Thomson.

Anexos

Anexo I Tablas y figuras referenciados en el informe

Anexo 1 Población Objetivo

Comunidad Autónoma	Tasa de variación 11/10 IPC	Tasa de variación 11/10 PIB	Tasa Paro	Tasa de variación 11/10 IPV
Andalucía	3.5	1.3	31.2	-6
Aragón	3	2.5	16.8	-8.2
Asturias	3.7	2.1	18.9	-5.7
Balears	2.7	3.1	25.2	-5.5
Canarias	2.5	3.4	30.9	-5.7
Cantabria	3.5	2.9	15.9	-9.3
Castilla y León	3.5	1.5	24.5	-6.2
Castilla-La Mancha	3.7	3.1	17.2	-6.4
Cataluña	3.3	2.6	20.5	-9.6
Comunidad Valenciana	3.3	1.8	25.4	-7.4
Extremadura	3.4	0.9	28.6	-6.1
Galicia	3.4	1.8	18.3	-5.1
Madrid	3.1	1.5	18.5	-8.7
Murcia	3.3	0.8	26.8	-5.1
Navarra	3	3.3	13.8	-7.7
País Vasco	3.1	3.3	12.6	-7.7
La Rioja	3.4	2.5	18.7	-11

Tabla A1.1 Datos macroeconómicos

Anexo 2 Metodología SEMMA

Estado civil	Frecuencia	Porcentaje
Soltero	316006	84.34
Casado	53620	14.31
Viudo	932	0.25
Separado	1489	0.4
Divorciado	2635	0.7

Estado Civil_a	Frecuencia	Porcentaje
Soltero	316006	84.34
Casado	53620	14.31
Separado/ Divorciado / Viudo	5056	1.35

Tabla A2.1 Recategorización Estado civil

Estructura hogar	Frecuencia	Porcentaje
Hogar con una mujer sola menor de 65 años	9503	2.54
Hogar con una hombre sola menor de 65 años	12188	3.25
Hogar con padre o madre que convive con algún hijo menor de 25 año	23478	6.27
Hogar con padre o madre que convive con todos sus hijos de 25 años o más	17659	4.71
Hogar formado por pareja sin hijos	42040	11.22
Hogar formado por pareja con hijos en donde algún hijo es menor de 25 años	139489	37.23
Hogar formado por pareja con hijos en donde todos los hijos de 25 años o más	60821	16.23
Hogar formado por pareja o padre/madre que convive con algún hijo menor de 25 años y otra(s) persona(a)	35227	9.4
Otro tipo de hogar	34277	9.15

Estructura hogar_a	Frecuencia	Porcentaje
Solos	21691	5.79
Monoparental	41137	10.98
Pareja sin hijos	42040	11.22
Pareja con hijos	200310	53.46
Otros	69504	18.55

Tabla A2.2 Recategorización Estructura del hogar

Año construcción vivienda	Frecuencia	Porcentaje	Porcentaje acumulado
antes 1900	15771	4.21	4.21
1900 a 1920	7469	1.99	6.2
1921 a 1940	9744	2.6	8.8
1941 a 1950	10805	2.88	11.69
1951 a 1960	23835	6.36	18.05
1961 a 1970	45573	12.16	30.21
1971 a 1980	80878	21.59	51.8
1981 a 1990	63118	16.85	68.64
1991 a 2001	52987	14.14	82.78
2002	4976	1.33	84.11
2003	5665	1.51	85.62
2004	6716	1.79	87.42
2005	9413	2.51	89.93
2006	9124	2.44	92.36
2007	8910	2.38	94.74
2008	8260	2.2	96.95
2009	5618	1.5	98.45
2010	3570	0.95	99.4
2011	2250	0.6	100

Año construcción vivienda_a	Frecuencia	Porcentaje	Porcentaje acumulado
Hasta 1950	43789	11.69	11.69
1951 a 1960	23835	6.36	18.05
1961 a 1970	45573	12.16	30.21
1971 a 1980	80878	21.59	51.8
1981 a 1990	63118	16.85	68.64
1991 a 2001	52987	14.14	82.78
2001 a 2011	64502	17.22	100

Tabla A2.3 Recategorización años de construcción de la vivienda

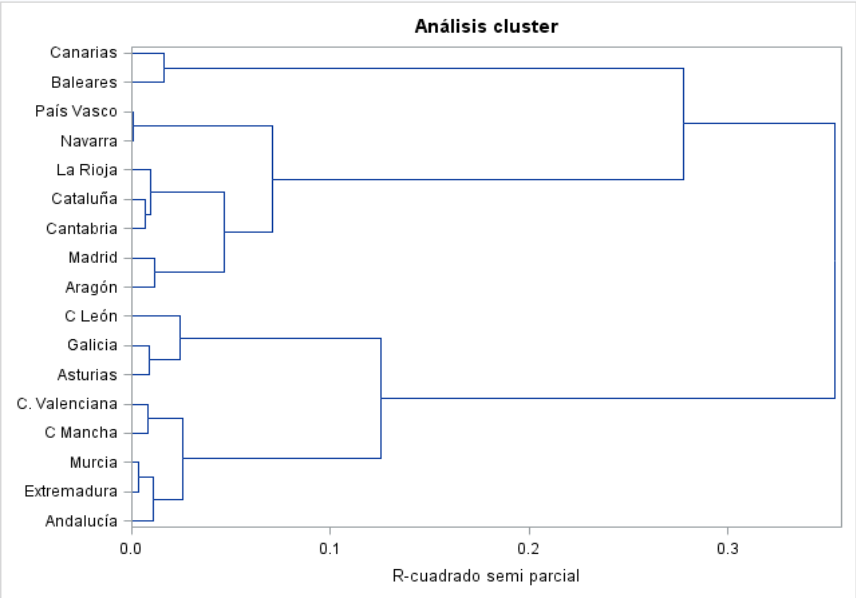


Figura A2.1 Dendograma

Contrastes multivariados[d]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.	Eta al cuadrado parcial	Parámetro de no centralidad Parámetro	Potencia observada[b]
Intersección	Traza de Pillai	.999	2982.899[a]	4.000	8.000	.000	.999	11931.595	1.000
	Lambda de Wilks	.001	2982.899[a]	4.000	8.000	.000	.999	11931.595	1.000
	Traza de Hotelling	1491.449	2982.899[a]	4.000	8.000	.000	.999	11931.595	1.000
	Raíz mayor de Roy	1491.449	2982.899[a]	4.000	8.000	.000	.999	11931.595	1.000
Clusters CCAA	Traza de Pillai	3.175	8.468	20.000	44.000	.000	.794	169.359	1.000
	Lambda de Wilks	.000	13.183	20.000	27.483	.000	.859	167.045	1.000
	Traza de Hotelling	33.431	10.865	20.000	26.000	.000	.893	217.299	1.000
	Raíz mayor de Roy	16.366	36.005[c]	5.000	11.000	.000	.942	180.023	1.000

a. Estadístico exacto
b. Calculado con alfa = .05
c. El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación.
d. Diseño: Intersección + CCAA

Tabla A2.4 Contrastes multivariados

Contraste de Levene sobre la igualdad de las varianzas error[a]

	F	gl1	gl2	Sig.
IPC_2011	1.379	5	11	.304
PAR_2011	3.521	5	11	.038
PIB_2011	3.562	5	11	.037
IPV_2011	1.022	5	11	.451

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección + CCAA

Tabla A2.5 Contraste Levene

Pruebas de los efectos inter-sujetos

Origen	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial	Parámetro de no centralidad Parámetro	Potencia observada[b]
Modelo corregido	IPC_2011	1.551[a]	5	.310	22.751	.000	.912	113.753	1.000
	PAR_2011	415.776[c]	5	83.155	16.190	.000	.880	80.950	1.000
	PIB_2011	9.771[d]	5	1.954	9.552	.001	.813	47.761	.994
	IPV_2011	43.511[e]	5	8.702	17.958	.000	.891	89.792	1.000
Intersección	IPC_2011	154.145	1	154.145	11303.972	.000	.999	11303.972	1.000
	PAR_2011	5046.726	1	5046.726	982.584	.000	.989	982.584	1.000
	PIB_2011	92.677	1	92.677	453.022	.000	.976	453.022	1.000
	IPV_2011	803.591	1	803.591	1658.339	.000	.993	1658.339	1.000
Clusters CCAA	IPC_2011	1.551	5	.310	22.751	.000	.912	113.753	1.000
	PAR_2011	415.776	5	83.155	16.190	.000	.880	80.950	1.000
	PIB_2011	9.771	5	1.954	9.552	.001	.813	47.761	.994
	IPV_2011	43.511	5	8.702	17.958	.000	.891	89.792	1.000
Error	IPC_2011	.150	11	.014					
	PAR_2011	56.498	11	5.136					
	PIB_2011	2.250	11	.205					
	IPV_2011	5.330	11	.485					
Total	IPC_2011	182.240	17						
	PAR_2011	6635.522	17						
	PIB_2011	98.760	17						
	IPV_2011	915.780	17						
Total corregida	IPC_2011	1.701	16						
	PAR_2011	472.274	16						
	PIB_2011	12.021	16						
	IPV_2011	48.841	16						

a. R cuadrado = .912 (R cuadrado corregida = .872)
b. Calculado con alfa = .05
c. R cuadrado = .880 (R cuadrado corregida = .826)
d. R cuadrado = .813 (R cuadrado corregida = .728)
e. R cuadrado = .891 (R cuadrado corregida = .841)

Tabla A2.6 Prueba de efectos inter-sujetos

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
IPC_2011	.197	17	.077	.926	17	.186
PAR_2011	.198	17	.075	.928	17	.199
PIB_2011	.139	17	.200*	.929	17	.207
IPV_2011	.194	17	.090	.917	17	.134

a. Corrección de la significación de Lilliefors
*. Este es un límite inferior de la significación verdadera.

Tabla A2.7 Prueba de normalidad

Anexo 3 Determinación de nivel estudios de los jóvenes

Estudios padre	Frecuencia	Porcentaje
Sin estudios	14121	7.8
Primarios	32649	18.03
Secundaria	107849	59.57
Universitarios	26418	14.59

Estudios Madre	Frecuencia	Porcentaje
Sin estudios	12901	7.13
Primarios	34164	18.87
Secundaria	111227	61.44
Universitarios	22745	12.56

Actividad padre	Frecuencia	Porcentaje
Empresario	16724	9.24
Antónomo	20464	11.3
Trabajador	96466	53.29
No aplicable	47383	26.17

Actividad madre	Frecuencia	Porcentaje
Empresario	6310	3.49
Antónomo	12441	6.87
Trabajador	96059	53.06
No aplicable	66227	36.58

Efecto calendario	Frecuencia	Porcentaje
1ª semestre	90969	50.25
2º semestre	90068	49.75

Vivienda	Frecuencia	Porcentaje
Otra situación	83113	45.91
Pagada	97924	54.09

Sexo	Frecuencia	Porcentaje
Hombre	99874	55.17
Mujer	81163	44.83

Hermanos	Frecuencia	Porcentaje
0 hermanos	65163	35.99
Tener 1 o más hermanos	115874	64.01

Comunidad Autónoma	Frecuencia	Porcentaje
País Vasco, Navarra	11952	6.6
Andalucía, Extremadura,Murcia, C. Valenciana, y C. la Mancha	73545	40.62
Cantabria, Cataluña, y La Rioja	26492	14.63
C. León, Asturias y Galicia	31732	17.53
Aragón y Madrid	29345	16.21
Canarias y Baleareas	7971	4.4

Área	Frecuencia	Porcentaje
< 20.000 Hab	88002	48.61
>20.000 Hab	93035	51.39

Nacionalidad_Padres	Frecuencia	Porcentaje
Española	175786	97.1
Alguno extranjero	2066	1.14
Ambos extranjeros	3185	1.76

Tabla A3.1 Análisis descriptivos de las variables

Estadístico	DF	Valor	Prob
Chi-cuadrado	6	16442.4351	<.0001
Chi-cuadrado de ratio de verosimilitud	6	16718.7461	<.0001
Chi-cuadrado Mantel-Haenszel	1	14296.47	<.0001
Coeficiente Phi		0.3014	
Coeficiente de contingencia		0.2886	
V de Cramer		0.2131	

Tabla A3.2 Test independencia estudios madre frente estudios hijos

Estadístico	DF	Valor	Prob
Chi-cuadrado	6	17423.5608	<.0001
Chi-cuadrado de ratio de verosimilitud	6	17785.8184	<.0001
Chi-cuadrado Mantel-Haenszel	1	14967.2	<.0001
Coeficiente Phi		0.3102	
Coeficiente de contingencia		0.2963	
V de Cramer		0.2194	

Tabla A3.3 Test independencia estudios padre frente estudios hijos

Tipo 3 Análisis de efectos			
Efecto	**DF**	**Chi-cuadrado**	**Pr > ChiSq**
Sexo	2	11559.039	<.0001
Área	1	223.223	<.0001
Vivienda	1	1211.742	<.0001
Comunidad Autónoma	10	895.467	<.0001
Efecto Calendario	2	14.344	0.001
Hermanos	2	64.093	<.0001
Nacionalidad_Padres	4	432.180	<.0001
Estudios padre	6	5161.367	<.0001
Estudios madre	6	4261.399	<.0001
Situación padre	6	399.191	<.0001
Situación madre	6	115.579	<.0001

Tabla A3.4 Significatividad de las variables test Wald

Anexo 4 Bloque Ocupación jóvenes universitarios

		Estadísticos de sumarización para los puntos			Contribuciones parciales a la inercia para los puntos		Cosenos cuadrados para los puntos	
		Calidad	**Mass**	**Inercia**	**Dim1**	**Dim2**	**Dim1**	**Dim2**
Tipo de estudios	Educación	0.8704	0.1753	0.0811	0.0929	0.033	0.783	0.0874
	Artes y Humanidades	0.6783	0.0741	0.028	0.0235	0.0126	0.4418	0.2365
	Derecho y Ciencias Sociales	0.9733	0.2692	0.3189	0.4217	0.0528	0.9377	0.0356
	Ciencias	0.9732	0.1195	0.0848	0.0188	0.324	0.1519	0.8213
	Arquitectura, Construcción, Formación Técnica e Industrias	0.9527	0.1239	0.0859	0.0307	0.2833	0.2443	0.7084
	Agricultura, Ganadería, Pesca y Veterinaria	0.564	0.0205	0.0453	0.004	0.046	0.0602	0.5038
	Salud y Servicios Sociales	0.9862	0.1723	0.2346	0.2471	0.2481	0.7196	0.2666
	Otros servicios	0.9085	0.0451	0.1214	0.1613	0.0002	0.908	0.0004
Ocupación	Directores y gerentes	0.8754	0.0254	0.0353	0.045	0.0008	0.8706	0.0047
	Técnicos y profesionales científicos e intelectuales	0.9899	0.5137	0.2451	0.312	0.1181	0.8863	0.1035
	Técnicos; profesionales de apoyo	0.986	0.1348	0.1637	0.000	0.7511	0.0002	0.9858
	Empleados contables, administrativos y otros empleados de	0.9907	0.1534	0.433	0.602	0.0602	0.9609	0.0299
	Trabajadores de los servicios de restauración, personales, protección y vendedores	0.5241	0.105	0.0542	0.031	0.0116	0.3981	0.126
	Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero	0.496	0.0057	0.0408	0.003	0.006	0.124	0.372
	Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción	0.6825	0.0201	0.0121	0.001	0.0351	0.0567	0.6258
	Operadores de instalaciones y maquinaria, y montadores	0.6359	0.0118	0.0065	0.0012	0.0142	0.1537	0.4822
	Ocupaciones elementales	0.6023	0.0299	0.0093	0.0048	0.0029	0.4071	0.1952

Tabla A4.1 Estadísticos relacionados con el análisis

Anexo 5 Bloque Factores emancipación

Comunidad Autónoma	Frecuencia	Porcentaje
Andalucía	66745	17.81
Aragón	14098	3.76
Asturias	6434	1.72
Balears	6391	1.71
Canarias	12640	3.37
Cantabria	4880	1.3
C. León	32804	8.76
C. La Mancha	24426	6.52
Cataluña	53015	14.15
C Valenciana	35127	9.38
Extremadura	13767	3.67
Galicia	19098	5.1
Madrid	46874	12.51
Murcia,	11052	2.95
Navarra	7304	1.95
País Vasco	16766	4.47
Rioja	3261	0.87

Nacionalidad	Frecuencia	Porcentaje
Española	342632	91.45
Extranjera	32050	8.55

Área	Frecuencia	Porcentaje
< 20.000 Hab	88002	48.61
>20.000 Hab	93035	51.39

Tabla A5.1 Estadísticos descriptivos

Tipo 3 Análisis de efectos			
Efecto	**DF**	**Chi-cuadrado de Wald**	**Pr > ChiSq**
Nacionalidad	1	4776.579	<.0001
Sexo	1	2473.6245	<.0001
Estado Civil	2	28868.3418	<.0001
Estudios	4	2022.8025	<.0001
Actividad	2	4589.5707	<.0001
Edad	1	21077.7592	<.0001
Área de residencia	1	25.5473	<.0001
CCAA	16	1673.7089	<.0001

Tabla A5.2 **Significatividad de las variables test Wald**

Anexo 6 Régimen de tenencia

Tenencia vivienda	Frecuencia	Porcentaje
Hipoteca	11442	62.97
Alquiler	6728	37.03

Área	Frecuencia	Porcentaje
< 20.000 Hab	88002	48.61
>20.000 Hab	93035	51.39

Nivel_Estudios	Frecuencia	Porcentaje
Primarios	1903	10.47
Ultimo curso ESO/Bachiller	5456	30.03
Bachiller	2435	13.4
Fp	4031	22.18
Universitarios	4345	23.91

Estructura hogar	Frecuencia	Porcentaje
Solos	2800	15.41
Monoparental	451	2.48
Pareja sin hijos	7171	39.47
Pareja con hijos	5208	28.66
Otros	2540	13.98

Sexo	Frecuencia	Porcentaje
Hombre	7394	40.69
Mujer	10776	59.31

Nacionalidad	Frecuencia	Porcentaje
Española	14755	81.21
Extranjera	3415	18.79

Actividad	Frecuencia	Porcentaje
Ocupado	11864	65.29
Parados	4692	25.82
Otra situación	1614	8.88

EstadoCiv	Frecuencia	Porcentaje
Soltero	10533	57.97
Casado	7251	39.91
Separado/ Divorciado / Viudo	386	2.12

C_Autonoma	Frecuencia	Porcentaje
País Vasco, Navarra	1179	6.49
Andalucía, Extremadura,Murcia, C. Valenciana, y C. la Mancha	7353	40.47
Cantabria, Cataluña, y La Rioja	3678	20.24
C. León, Asturias y Galicia	2097	11.54
Aragón y Madrid	3015	16.59
Canarias y Baleareas	848	4.67

Año construcción	Frecuencia	Porcentaje
Hasta 1960	2883	15.87
1961 a 1970	2034	11.19
1971 a 1980	2594	14.28
1981 a 1990	1357	7.47
1991 a 2001	2386	13.13
2001 a 2011	6916	38.06

Tabla A6.1 Análisis descriptivo

Análisis del estimador de máxima verosimilitud							
Parámetro	Categoría	Referencia	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept			1	8.0602	0.3882	431.1065	<.0001
Nivel_Estudios	Secundaria	Primarios	1	-0.2382	0.0341	48.8643	<.0001
	Bachiller		1	0.1479	0.0445	11.0384	0.0009
	Fp		1	-0.2448	0.0384	40.7138	<.0001
	Universitarios		1	0.4109	0.0381	116.086	<.0001
Sexo	Mujer	Hombre	1	-0.1089	0.0198	30.3764	<.0001
Aréa	> 20.000 Hab	< 20.000 Hab	1	0.1125	0.0205	30.0102	<.0001
Est_Hogar	Monoparental	Solos	1	-0.1304	0.0971	1.8039	0.1792
	Pareja sin hijos		1	-0.2136	0.0389	30.1892	<.0001
	Pareja con hijos		1	-0.3006	0.0506	35.3314	<.0001
	Otros		1	0.4017	0.0505	63.2192	<.0001
Actividad	Paro	Ocupado	1	0.0385	0.0435	0.7848	0.3757
	Otra Situación		1	-0.1186	0.0482	6.067	0.0138
EstadoCiv	Casado	Soltero	1	-0.4045	0.0493	67.3732	<.0001
	Separado		1	0.0842	0.0854	0.9707	0.3245
Nacionalidad	Extranjera	Española	1	0.9701	0.027	1287.938	<.0001
C Autónoma	Andalucía, Extremadura,Murcia, C. Valenciana, y C. la Mancha	País Vasco, Navarra	1	-0.61	0.0503	147.1413	<.0001
	Cantabria, Cataluña, y La Rioja		1	0.646	0.1008	41.1162	<.0001
	C. León, Asturias y Galicia		1	0.00515	0.0777	0.0044	0.9472
	Aragón y Madrid		1	0.3451	0.0676	26.0737	<.0001
	Canarias y Baleareas		1	-0.0933	0.0948	0.9682	0.3251
N_Habitaciones			1	-0.1999	0.0173	132.7312	<.0001
Ratio_Ocupados			1	-0.5751	0.0844	46.4466	<.0001
Superficie			1	-0.00725	0.000719	101.8581	<.0001
Año_Construcción			1	-0.1808	0.00999	327.5257	<.0001
Edad			1	-0.1545	0.00879	308.8814	<.0001
IPV_2011			1	-0.1699	0.039	19.0036	<.0001

Tabla A6.2 Análisis descriptivo

Anexo II Cuestionario *"Censo de Población y Viviendas 2011"*

Censos de Población y Viviendas 2011

Instituto Nacional de Estadística

¿Qué son los Censos?

Los Censos de Población y Viviendas sirven para conocer las características de la población y poder así planificar y organizar servicios públicos (construcción de hospitales, carreteras, políticas sociales...) y actividades privadas (instalación de supermercados, entidades bancarias...).
Se realizan en cumplimiento del Reglamento CE Nº 763/ 2008 del Parlamento Europeo y del Consejo de la Unión Europea.

Obligación de responder

Es obligatorio responder a los Censos de Población y Viviendas (Ley 13/1996). Si no responde o si se dan premeditadamente datos falsos, se podrán aplicar las sanciones previstas en los artículos 50 y 51 de la Ley 12/1989 de la Función Estadística Pública.

Secreto estadístico

La información que proporcione es confidencial y está protegida por el secreto estadístico (Ley 12/1989). En particular, **no será publicada ninguna información de manera que se pueda saber a quién corresponde, ni siquiera indirectamente.**

¿Cómo hay que responder?

Por favor, rellene este cuestionario, en los 15 días siguientes a su recepción:

▶ **Por Internet,** en la dirección: **www.censos2011.es**. En este caso necesitará las dos claves que figuran en la carta que se incluye en el sobre. Es muy fácil y cómodo, la aplicación es accesible y dispone de ayudas.

▶ **O por correo:** rellene este cuestionario y envíelo por correo en el sobre que se adjunta y que no necesita franqueo. Si elige esta opción, por favor, antes lea atentamente las instrucciones que hay a la derecha.

Por favor, facilítenos un número de teléfono y una persona de contacto a la que llamaremos si es necesario realizar alguna aclaración.

Teléfono 1: | | | | | | | | | |

Teléfono 2: | | | | | | | | | |

Nombre y apellidos (Ejemplo): JUAN LOPEZ GOMEZ

¿Cómo rellenar el cuestionario?

Si ha elegido responder por Internet no tiene que rellenar este cuestionario.

Para responder por correo siga estas instrucciones:

- Compruebe que ha sacado del sobre todo el material.
- Lea la carta de presentación si aún no lo ha hecho.
- Comience contestando el **Cuestionario de Vivienda** que está a la vuelta de esta hoja.
- A continuación encontrará los **Cuestionarios Individuales**. Deberá rellenar uno por cada persona que viva en esta vivienda.
- Si viven más de 6 personas en la vivienda llame al teléfono **900 XXX XXX**. La llamada es gratuita.
- Para las preguntas **15** y **16** necesitará las listas de ocupaciones y actividades que figuran en el folleto.
- Una vez haya rellenado los cuestionarios individuales para cada una de las personas, meta este cuadernillo en el sobre de envío gratuito y envíelo por correo.

Por favor, tenga en cuenta...

Este cuestionario será leído por un escáner.
Le pedimos que :

- Para responder, marque con un aspa (X) el cuadro que corresponda a su respuesta. Si se equivoca, tache completamente y marque la opción correcta: ■ 1 ☒ 2 ☐ 3
- Escriba con letras **MAYÚSCULAS** y sin acentos. Use **una casilla para cada letra** y separe las palabras con un espacio en blanco. Si se equivoca, tache completamente la casilla:

Municipio:

S	A	N		S	E	■	B	A	S	T	I	A	N		D	E		L	A		G	O	M
E	■	R	A																				

Use bolígrafo azul o negro
(nunca lápiz ni bolígrafo rojo)

¿Necesita ayuda?

Teléfono gratuito: 900 XXX XXX

www.censos2011.es

Cuestionario de Vivienda

1. Lista de personas

Escriba el nombre y los apellidos de cada una de las personas que viven habitualmente en esta vivienda.

Debe incluir:

- a todas las personas que viven en esta vivienda la mayor parte del año, aunque no tengan lazos familiares
- a los hijos/as estudiantes que están ausentes durante el curso académico
- a los hijos/as en custodia compartida si viven en esta vivienda la mayor parte del tiempo

No olvide incluir:

- a los niños/as pequeños o recién nacidos
- ni a usted mismo si vive aquí

Nombre y apellidos (Ejemplo): JUAN LOPEZ GOMEZ

Persona nº 1

Persona nº 2

Persona nº 3

Persona nº 4

Persona nº 5

Persona nº 6

¡RECUERDE! Si viven más de 6 personas en esta vivienda llame al teléfono gratuito 900 XXX XXX

2. Propiedad de la vivienda

La vivienda es...

- ☐ Propia, por compra, totalmente pagada
- ☐ Propia, por compra, con pagos pendientes (hipotecas...)
- ☐ Propia por herencia o donación
- ☐ Alquilada
- ☐ Cedida gratis o a bajo precio (por otro hogar, pagada por la empresa...)
- ☐ Otra forma

3. ¿Cuáles de las siguientes instalaciones tiene la vivienda?

Calefacción

- ☐ Colectiva o central
- ☐ Individual
- ☐ NO tiene instalación de calefacción pero sí algún aparato que permite calentar alguna habitación (ejemplo: radiadores eléctricos)
- ☐ NO tiene calefacción

Cuarto de aseo con inodoro (WC, retrete)

☐ SÍ ☐ NO

Ducha o bañera

☐ SÍ ☐ NO

4. ¿Tiene la vivienda contratado servicio de acceso a Internet?

☐ SÍ ☐ NO

5. ¿Cuál es el sistema de suministro de agua?

- ☐ Agua corriente por abastecimiento público
- ☐ Agua corriente por abastecimiento privado o particular del edificio
- ☐ No tiene agua corriente

6. ¿Cuál es aproximadamente la superficie útil de la vivienda?

No incluya espacios que no sean habitables como terrazas abiertas o jardines; tampoco sótanos, desvanes, trasteros...

☐☐☐ m^2

7. ¿Cuántas habitaciones tiene la vivienda?

Incluya la cocina, los dormitorios y todas las habitaciones que tengan 4 metros cuadrados o más.
NO incluya cuartos de baño, vestíbulos, pasillos, terrazas abiertas...

☐☐ habitaciones

¡ATENCIÓN!

Pase a rellenar un Cuestionario Individual para cada una de las personas que ha incluido en la Lista de personas de arriba. Por favor, asegúrese de seguir el mismo orden que en la Lista de personas.

Cuestionario Individual de la Persona 1

Escriba los siguientes datos para la persona que aparece en primer lugar (Persona nº 1) en la **Lista de personas**.

Nombre y apellidos:

Fecha de nacimiento: día mes año

Sexo: ☐ Hombre ☐ Mujer

País de nacimiento:

☐ España. *Escriba municipio y provincia:*

Municipio:

Provincia:

☐ Otro país:

¿Cuál es su nacionalidad?

Si tiene doble nacionalidad, española y otra, marque ambas opciones y escriba el país correspondiente.

Si tiene doble nacionalidad, pero ninguna es la española, escriba únicamente una de ellas.

☐ Española

☐ De otro país:

¡ATENCIÓN! Conteste a las preguntas en orden y siguiendo las indicaciones

1 ¿Desde qué año reside en esta vivienda?

Desde el año ☐ Desde que nació (Pase a la pregunta 3)

¿y en este municipio?

Desde el año ☐ Desde que nació (Pase a la pregunta 3)

¿y en esta comunidad autónoma?

Desde el año ☐ Desde que nació

¿y en España?

Desde el año ☐ Desde que nació

¿Dónde residía antes de llegar por última vez a este municipio?

☐ En otro municipio:

Provincia:

☐ En otro país:

2 ¿Dónde residía hace 1 año?

☐ En este municipio (o no había nacido aún)

☐ En otro municipio:

Provincia:

☐ En otro país:

¿Dónde residía hace 10 años?

☐ En este municipio (o no había nacido aún)

☐ En otro municipio:

Provincia:

☐ En otro país:

3

Espacio reservado para preguntas específicas sobre la lengua en Comunidades Autónomas con lengua cooficial diferente al castellano.

(Pase a la pregunta 4)

4 En los últimos doce meses, ¿ha pasado más de 14 noches *(aunque no fueran seguidas)* **en otro municipio de España o en otro país?**

Puede ser por razones de trabajo, estudio, fines de semana, vacaciones o porque reside en más de un municipio

☐ NO → (Pase a la pregunta 5)

☐ SÍ → *Indique el lugar donde ha pasado más noches y el número aproximado de noches que ha pasado allí:*

☐ Otro municipio:

Provincia:

☐ Otro país:

Nº de noches:

¿Dispone en este lugar de una segunda vivienda (ya sea en propiedad, alquiler o cedida gratis)?

☐ SÍ ☐ NO

Cuestionario Individual de la Persona 1 (continuación)

5 ¿Cuál es el estado civil de la Persona 1?

Marque su situación legal, aunque no coincida con la de hecho

☐ Soltero/a ☐ Separado/a ☐ Viudo/a
☐ Casado/a ☐ Divorciado/a

6 ¿Viven en esta vivienda los siguientes familiares de la Persona 1?

*Cuando marque SI, escriba el número con el que este familiar aparece en la **Lista de personas** de la página 2:*

Padre ☐ NO ☐ SÍ → Persona nº ☐
Madre ☐ NO ☐ SÍ → Persona nº ☐
Cónyuge o pareja ☐ NO ☐ SÍ → Persona nº ☐
Otros familiares (hijos, hermanos, etc.) ☐ NO ☐ SÍ → Nº de persona de uno de ellos ☐

7 ¿Dónde nació el padre de la Persona 1?

Responda sólo si el padre no vive en esta vivienda:

☐ España.
☐ Otro país:

8 ¿Dónde nació la madre de la Persona 1?

Responda sólo si la madre no vive en esta vivienda:

☐ España.
☐ Otro país:

Si la Persona 1 tiene 16 años o más, pase a la pregunta 10.
Si tiene menos de 16 años, continúe en la pregunta 9.

9 ¿Va a algún centro escolar *(incluyendo guarderías)*?

☐ SÍ → (Pase a la pregunta 19)
☐ NO → ***Ya ha terminado el Cuestionario Individual de la Persona 1***

10 ¿Cuál es el mayor nivel de estudios que ha completado?

Marque sólo una opción

☐ No sabe leer o escribir
☐ Sabe leer y escribir pero fue menos de 5 años a la escuela
☐ Fue a la escuela 5 o más años pero no llegó al último curso de ESO, EGB o Bachiller Elemental
☐ Llegó al último curso de ESO, EGB o Bachiller Elemental o tiene el Certificado de Escolaridad o de Estudios Primarios
☐ Bachiller (LOE, LOGSE), BUP, Bachiller Superior, COU, PREU

(Pase a la pregunta 12)

☐ FP grado medio, FP I, Oficialía Industrial o equivalente, Grado Medio de Música y Danza, Certificados de Escuelas Oficiales de Idiomas
☐ FP grado superior, FP II, Maestría Industrial o equivalente
☐ Diplomatura universitaria, Arquitectura Técnica, Ingeniería Técnica o equivalente
☐ Grado Universitario o equivalente
☐ Licenciatura, Arquitectura, Ingeniería o equivalente
☐ Máster oficial universitario (a partir de 2006), Especialidades Médicas o análogos
☐ Doctorado

11 ¿A qué campo corresponden esos estudios?

Marque sólo una opción y escriba en la fila de cuadros el nombre de la titulación obtenida

☐ Educación (Magisterio, Educación Infantil, Pedagogía...)
☐ Artes y Humanidades (Historia, Lenguas, Imagen y Sonido...)
☐ Derecho y Ciencias Sociales (Administración, Psicología, Economía, Periodismo...)
☐ Ciencias (Biología, Química, Física, Matemáticas...) e Informática (incluida Ingeniería Informática)
☐ Arquitectura, Construcción, Formación Técnica e Industrias (Mecánica, Metal, Electrónica, Diseño, Confección, Alimentación, etc., incluidas Ingenierías de estos campos)
☐ Agricultura, Ganadería, Pesca y Veterinaria (incluida Ingeniería Agrónoma o similar)
☐ Salud y Servicios Sociales (Medicina, Enfermería, Farmacia, Trabajo Social...)
☐ Otros servicios (Turismo, Hostelería, Peluquería, Enseñanza Náutica, Militar...)

Escriba el nombre de la titulación obtenida:

12 Durante la semana pasada, ¿realizó alguna de las siguientes tareas sin recibir dinero a cambio?

Puede marcar varias opciones

☐ Cuidar a un/una menor de 15 años
☐ Cuidar a una persona con problemas importantes de salud
☐ Tareas benéficas o de voluntariado social
☐ Encargarse de la mayor parte de las tareas domésticas de su hogar
☐ No realizó ninguna de esas tareas

13 Si la Persona 1 es mujer, ¿ha tenido hijos/as?

Considere todos los que nacieron con vida

☐ NO ☐ SÍ → ¿Cuántos? ☐☐

14 ¿En qué situación laboral estaba la semana pasada?

☐ Ocupado/a (es decir, trabajó al menos una hora) o temporalmente ausente del trabajo:
  ☐ a tiempo completo ☐ a tiempo parcial
☐ Parado/a que ha trabajado antes
☐ Parado/a buscando su primer empleo
☐ Persona con invalidez laboral permanente
☐ Jubilado/a, prejubilado/a, pensionista o rentista
☐ Otra situación

(Pase a la pregunta 18)

Cuestionario Individual de la Persona 1 (continuación)

15 ¿Cuál es (era) su ocupación?

Si está parado refiérase a su última ocupación

Búsquela en la **TABLA DE OCUPACIONES** *del* **FOLLETO ADJUNTO** *y anote la letra y el número que le corresponde:*

☐ Letra ☐ Número

Si no ha encontrado su ocupación o tiene dudas, escríbala a continuación:

16 ¿Cuál es (era) la actividad principal del establecimiento o local donde trabaja (trabajaba)?

Si está parado refiérase a su última ocupación

Búsquela en la **TABLA DE ACTIVIDADES** *del* **FOLLETO ADJUNTO** *y anote el número que le corresponde:*

Número ☐☐☐

Si no ha encontrado su actividad o tiene dudas, escríbala a continuación:

17 ¿Cuál es (era) su situación profesional?

Marque sólo una opción

Empresario, profesional o trabajador por cuenta propia:

☐ que emplea personal ☐ que no emplea personal

Trabajador por cuenta ajena:

☐ fijo o indefinido ☐ eventual o temporal

Otras situaciones:

☐ ayuda familiar ☐ miembro de cooperativas

18 ¿Está recibiendo algún tipo de enseñanza *(incluso en academias, cursos organizados por su empresa...)?*

☐ NO → ¿Trabajó al menos 1 hora la semana pasada?

SÍ → (Pase a la pregunta 19)

NO → ***Ya ha terminado el Cuestionario Individual de la Persona 1***

☐ SÍ → *Señale cuál o cuáles hasta un máximo de 3:*

- ☐ ESO, Educación secundaria para adultos
- ☐ Programas de Cualificación Profesional Inicial
- ☐ Bachillerato
- ☐ Grado Medio de FP, de Artes Plásticas y Diseño y de Enseñanzas Deportivas o equivalente
- ☐ Enseñanzas de Escuelas Oficiales de Idiomas
- ☐ Enseñanzas Profesionales de Música y Danza
- ☐ Grado Superior de FP, de Artes Plásticas y Diseño y de Enseñanzas Deportivas o equivalente
- ☐ Diplomatura universitaria, Arquitectura Técnica, Ingeniería Técnica o equivalente
- ☐ Estudios de Grado Universitarios y de Enseñanzas Artísticas y equivalentes
- ☐ Licenciatura, Arquitectura, Ingeniería o equivalente
- ☐ Máster oficial universitario, Especialidades Médicas o análogos
- ☐ Doctorado
- ☐ Otros cursos de educación reglada (Enseñanzas iniciales para adultos...)
- ☐ Cursos de formación de los Servicios Públicos de Empleo
- ☐ Otros cursos de formación no reglados

19 ¿Dónde está su lugar de trabajo o de estudio?

Si trabaja y estudia, le preguntamos por el sitio donde trabajó la semana pasada

☐ En el propio domicilio

☐ En varios municipios (comercial, conductor...)

→ ***Ya ha terminado el Cuestionario Individual de la Persona 1***

☐ En este municipio

☐ En otro municipio:

Provincia:

☐ En otro país:

Código postal de su centro de trabajo o de estudios:

20 ¿Cuántos viajes de ida y vuelta realiza normalmente cada día desde esta vivienda hasta el lugar de trabajo o estudio?

☐ Ninguno (porque tiene una segunda residencia desde la que se desplaza a trabajar o estudiar)

→ ***Ya ha terminado el Cuestionario Individual de la Persona 1***

☐ Uno diario (es decir, uno de ida y otro de vuelta)

☐ Dos o más diarios

21 ¿Cómo va normalmente desde su casa hasta el lugar de trabajo o estudio?

Si utiliza varios medios de transporte, marque los dos con los que cubre más distancia

- ☐ En coche o furgoneta como conductor
- ☐ En coche o furgoneta como pasajero
- ☐ En autobús, autocar, minibús
- ☐ En metro
- ☐ En moto
- ☐ Andando
- ☐ En tren
- ☐ En bicicleta
- ☐ Otros medios

22 ¿Cuánto tiempo tarda normalmente desde su casa hasta el lugar de trabajo o estudio?

- ☐ Menos de 10 minutos
- ☐ Entre 10 y 19 minutos
- ☐ Entre 20 y 29 minutos
- ☐ Entre 30 y 44 minutos
- ☐ Entre 45 y 59 minutos
- ☐ Entre 1 hora y hora y media
- ☐ Más de hora y media

Ya ha terminado el Cuestionario Individual de la Persona 1. Continúe con la Persona 2 o, si no hay más personas, ya ha terminado de rellenar el censo. Meta este cuadernillo en el sobre de envío gratuito y envíelo por correo.

¡MUCHAS GRACIAS POR SU COLABORACIÓN!

00000001 03