Facultad de Ciencias

APLICACIÓN DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO AL FENÓMENO DE LA DESPOBLACIÓN

Application of Machine Learning Techniques to the depopulation phenomenom.

Trabajo de Fin de Máster
para acceder al

**MÁSTER EN CIENCA DE DATOS /
MASTER IN DATA SCIENCE**

Autor: Laura Muñiz Jaén

Director\es: Francisco Matorras Weinig

Olga de Cos Guerra

Resumen

El presente trabajo pretende abordar un problema de carácter demográfico desde un punto de vista novedoso aplicando técnicas de aprendizaje automático, en concreto, redes neuronales. Se busca poder predecir el fenómeno de la despoblación en los municipios en un periodo de 10 años a partir de los datos obtenidos por los censos que se realizan sobre el territorio nacional.

Basándose en el modelo que resultó tener un buen valor predictive de la despoblación se estudiará la dependencia de las variables con este fenómeno. En base a este modelo, se han identificado variables sobre las que se podría actuar, que revertirían la evolución de la población, proporcionando una herramienta para evitar este fenómeno.

Abstract

This works pretends to solve a demographic problem from a new point of view using machine learning techniques, in particular, neural networks. The aim of this study is to predict the depopulation phenomenon municipalities in a 10 years period using the data obtained from census with information about the national territory.

Moreover, once an optimum model has been obtained the dependence between the depopulation prediction an the different variables is going to be studied and how affect to the depopulation trend changes in this variables.

Palabras clave: Despoblación, Aprendizaje Automático, Redes Neuronales.

INDICE

1 Introducción **4**
1.1 Marco teórico. Despoblación rural y envejecimiento 4
1.2 Aprendizaje Automático 5
1.3 Redes neuronales 6
1.3.1 Overfitting y Underfitting 8
1.4 Redes Neuronales en R 9

2 Datos **10**
2.1 Descripción de los datos 10
2.2 Carga de datos en R 11
2.3 Curación de datos 11

3 Análisis preliminar de las variables **15**
3.1 Variables territoriales 15
3.2 Variables del censo de 2001 16
3.3 Poder discriminante de cada variable 15

4 Diseño de una Red Neuroanl **21**
4.1 Construcción de la red 23
4.2 Ajuste de parámetros 16

5 Clasificación **26**
5.1 Primer modelo 26
5.2 Segundo modelo 27
5.3 Reducción de variables 29
5.4 Comparativa de los modelos 33

6 Análisis para la NN seleccionada **34**
6.1 Variables con mayor dependencia con la red 34

7 Estudio de la posibilidad de revertir la tendencia poblacional 40

8 Conclusiones 48

Bibliografía 49
Anexo 50

1. Introducción

1.1 Marco teórico. Despoblación rural y envejecimiento.

Actualmente existe una creciente preocupación por procesos como la despoblación y el envejecimiento de la población, fenómenos que se superponen en España, centrándose en las zonas rurales, y que están ganando importancia en las agendas políticas.

Existe un destacado interés por detectar y analizar las áreas más vulnerables en distintos ámbitos y escalas y desde diferentes dimensiones: socio-demográfica, socio-económica, residencial, percibida o subjetiva [3]. Las áreas de elevada concentración de habitantes ya han sido foco de estudio anteriormente, pero en la actualidad la preocupación se centra en las situaciones de vulnerabilidad derivadas de niveles de ocupación tan bajos que los territorios están entrando en una trayectoria crítica desde el punto de vista de la sostenibilidad demográfica y territorial, presente y futura.

Estas variables demográficas están ligadas a los factores territoriales, económicos, sociales o políticos entre otros, por lo que su estudio no debería de ser aislado, sino realizado de manera transversal entre las distintas disciplinas.

La población es el elemento básico de las estructuras territoriales y económicas y de las organizaciones sociales. Este elemento ha presentado a lo largo de su historia un carácter altamente dependiente de conceptos como la industrialización, el crecimiento urbano, infraestructuras de transporte…, es decir, unos factores tanto económico-territoriales como político-territoriales. [3]

El contexto en el territorio español en este tema es más crítico con respecto a otros países debido a que el proceso de cambio y modernización de las estructuras agrarias y, como consecuencia, el éxodo rural, se ha desarrollado más tardíamente, pero de manera más intensa y en un periodo de tiempo menor, en tan solo dos o tres décadas.

Estas transformaciones han derivado en una pérdida de población masiva en el entorno rural que impide asegurar el reemplazo generacional. A este hecho se le suma el envejecimiento del empresario del sector primario y las densidades de población por debajo del umbral crítico.

Por tanto, la despoblación, y sus consecuencias económicas, en la mayor parte de las zonas rurales y singularmente en las zonas de montaña, se han convertido en uno de los temas emergentes tanto político como sociales de la mayor parte de los países europeos. En especial de España donde en 2015 se creó la Comisión de Despoblación en la Federación Española de Municipios y Provincias [5] y en 2017 el Gobierno de España creó el Comisionado frente al reto Demográfico [6].

Sobre esta base teórica y este contexto actual se pretende realizar un estudio de la despoblación desde un punto de vista novedoso. En vez de abordar el problema desde un punto de vista teórico más centrado en conceptos geográficos y demográficos se quiere intentar transformar en un problema de aprendizaje automático. Es decir, no se va a profundizar en un análisis sobre cada municipio ni sobre el significado de cada variable,

sino que se va a intentar abordar el problema como un problema de clasificación de aprendizaje automático.

El objetivo es poder predecir la despoblación de un municipio en un periodo de 10 años, que es el rango de tiempo que existe entre la elaboración de los censos. La información disponible procede de las fuentes censales de los años 2001 y de 2011 además de una serie de variables territoriales de los municipios.

A parte de predecir la despoblación otro objetivo es intentar buscar la relación de las variables con este fenómeno para intentar actuar sobre ellas.

Se pretende que este estudio permita determinar si es posible abordar problemas en este campo desde un punto de vista más centrado en la ciencia de datos así como poder saber que variables son las más relevantes para poder actuar sobre ellas o para la futura elaboración censal.

Además, desde un punto de vista demográfico se va a realizar un estudio del fenómeno de la despoblación en el territorio nacional con un método que integra casuísticas autonómicas y territoriales muy distintas y que considera conjuntamente las variables de tipo territorial y demográficas, siendo estas últimas las más habituales en este tipo de estudios.

1.2 Aprendizaje automático

El término Machine Learning se refiere a la detección automática de patrones en los datos con un significado en. Debido a los grandes volúmenes de datos que se generan, y, por tanto, con los que se trabaja actualmente se ha convertido en una técnica muy utilizada en casi cualquier campo de estudio. Las distintas tareas que engloba el Machine Learning son tareas de reconocimiento diagnóstico, predicción..

El Machine Learning se puede dividir en dos subcategorías dependiendo del tipo de aprendizaje de sus algoritmos, supervisado o no supervisado.

Es el primer tipo en el que se va a basar el posterior modelo desarrollado para la predicción de la despoblación, por lo que se explicará con más detalle.

Los algoritmos de aprendizaje supervisado desarrollan la tarea de predecir una variable objetivo a partir de una serie de inputs. La variable objetivo o etiqueta es normalmente denominada como ***y***. Los datos de entrada, o variables predictoras se denominan como ***x***.

Se puede definir un dataset como una colección de *n* instancias $\{x_i,y_i\}^n{}_{i=1.}$
El objetivo del aprendizaje supervisado es producir un modelo f_ϑ que mapee el input de $\boldsymbol{x_i}$ en una predicción $f_\vartheta(x_i)$.

La parte “supervisada” de este aprendizaje viene cuando para elegir los parámetros ϑ, se proporciona al modelo un dataset con las variables predictoras y la variable objetivo que

debe de dar, es decir, se proporciona un dataset ya etiquetado con ejemplos de los inputs y su etiqueta correspondiente.

El proceso de aprendizaje es el siguiente, el modelo se alimenta a partir de los inputs y las etiquetas de aprendizaje para posteriormente poder predecir a partir de unos inputs dados.

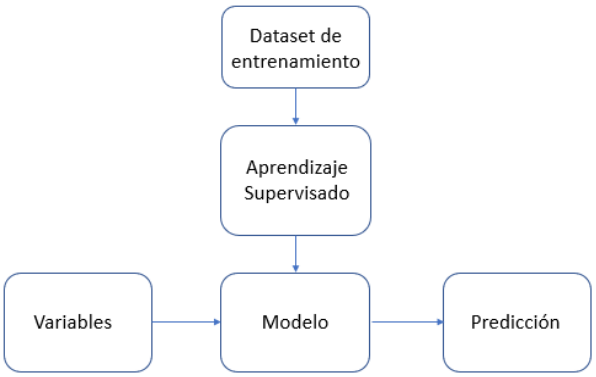


Figura 1. Esquema del proceso de un modelo de aprendizaje automático con aprendizaje supervisado. Elaboración propia.

1.3 Redes Neuronales

A pesar de que las redes neuronales fueron propuestas en el año 1950 fue tiempo después cuando empezaron a ganar importancia debido al avance computacional [1]. Actualmente tienen un uso destacado y han probado dar resultados fiables para problemas que usando otras técnicas computacionales eran difíciles o imposibles de resolver.

Las redes neuronales son modelos computacionales que se inspiran en las características neurológicas. Las neuronas biológicas constan de tres partes: el cuerpo de la neurona, las dendritas (que reciben la información de entrada) y el axón (que lleva la salida de la neurona a las dendritas de otras). Aunque la forma exacta de cómo se relacionan no se conoce, de forma general la neurona a través de su axón envía la información de salida a otras en forma de diferencia de potencial eléctrico. La neurona recoge todas las señales, que pueden ser excitadoras, positivas, o inhibidoras, negativas. Dependiendo de qué tipo domine manda una señal positiva o negativa.

Una red neuronal computacional consiste en nodos, o neuronas, que están interconectados. La mayoría de las redes organizan estos nodos en capas y envían la información en una única dirección, “feed-forward”. Una neurona puede estar conectada a distintas neuronas de la capa anterior, de las cuales recibirá los datos, y a diferentes nodos de la capa siguiente, a los que enviará la información.

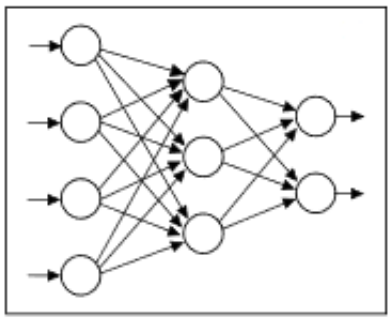


Figura 2. Esquema sobre la conexión entre las neuronas de distintas capas en una red neuronal. [1]

A cada una de las conexiones de entrada que tiene una neurona se le asigna un peso determinado por el que se multiplica su valor de salida. La suma de los productos de las distintas conexiones de entrada que tiene la neurona es lo que se conoce como actividad lineal de la neurona, ***Yi.***

$$Y_i = \sum_{j=1}^{n} w_{ij} x_j$$

Siendo w_{ij} el peso que se asigna a una determinada conexión y x_j el valor de salida de esa neurona.

El valor de salida de una neurona $\boldsymbol{y_i}$ se obtiene transformando la actividad linear con una función de activación:

$$y_i = f(\sum_{j=1}^{n} w_{ij} x_j)$$

Hay distintos tipos de función de activación, entre ellos destacan: activación lineal, funciones a pasos o la función sigmoidal.

La función de activación sigmoide es una función que da una salida no linear que se define como:

$$f_c = \frac{1}{1 + e^{-cx}}$$

El output de la neurona se obtiene usando la función de activación y se buscan funciones de activación simples para reducir el coste computacional.

Trazando el valor que devuelve la función sigma con $\boldsymbol{x}$ como entrada se obtiene

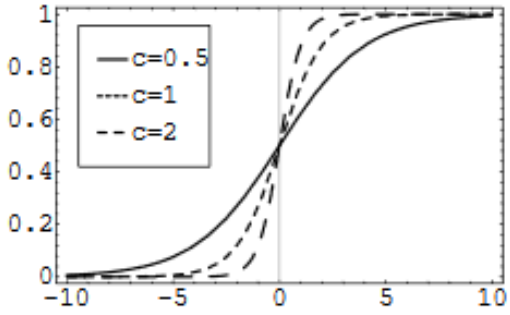


Figura 3. Función de activación sigmoide. [1]

Como se puede observar en la **Figura 3** la función transforma el input en una salida en un rango de 0 a 1, para valores negativos de ***x*** la función se aproxima a cero y para valores positivos tiende a 1. Esta función es usada principalmente en tareas de clasificación donde se tiene que predecir la probabilidad de la salida, ya que la probabilidad es un valor que se encuentra siempre entre 0 y 1.

Los datos de entrenamiento alimentan la capa de entrada pasando a través de las siguientes capas siendo transformados hasta que llegan a la capa de salida. El valor del peso de las distintas conexiones se asigna inicialmente como un valor aleatorio y durante el entrenamiento es continuamente ajustado hasta conseguir que los datos con las mismas etiquetas proporcionen resultados similares de forma consistente.

En el aprendizaje supervisado se conocen los patrones de entrada junto con su salida correspondiente. De este modo cada neurona de salida conoce el output deseado para un determinado input. La forma de obtener los pesos es normalmente minimizando la función de error que mide la diferencia entre lo deseado y lo obtenido. Un problema que surge con este tipo de aprendizaje es la convergencia del error, ya que en muchos casos al minimizar la función de error se puede encontrar un mínimo local sin obtener el mínimo global de la función.

Dentro de la Geografía ha surgido una subdisciplina cuyo objetivo es otorgar un mayor protagonismo al uso de la informática para la explicación y resolución de problemas geográficos, se conocen como Geocomputación y es una disciplina transdisciplinar formada por la Geografía, la Informática y la Inteligencia Artificial. Dentro de este campo las redes neuronales son una herramienta útil para el modelado de sistemas espaciales que puede sustituir a los métodos estadísticos convencionales que son afectados por problemas como la autocorrelación espacial. [13]

1.3.1 Overfitting y Underfitting.

Cuando un modelo de redes neuronales durante su entrenamiento comienza a perder la capacidad de mejorar su habilidad de resolver el problema se denomina que existe overfitting. Este es uno de los mayores y más habituales problemas que surgen al implementar redes neuronales. La red comienza a aprender patrones aleatorios contenidos en el conjunto de entrenamiento y pierde la capacidad de generalización del modelo.

Otro problema que surge, opuesto al overfitting pero al que también hay que tener en cuenta durante el entrenamiento de las redes neuronales, es el underfitting. En este caso el modelo no consigue aprender a reconocer los patrones de los datos de entrenamiento y no aprende el patrón principal. Este fenómeno suele ocurrir cuando el modelo no ha tenido suficiente entrenamiento.

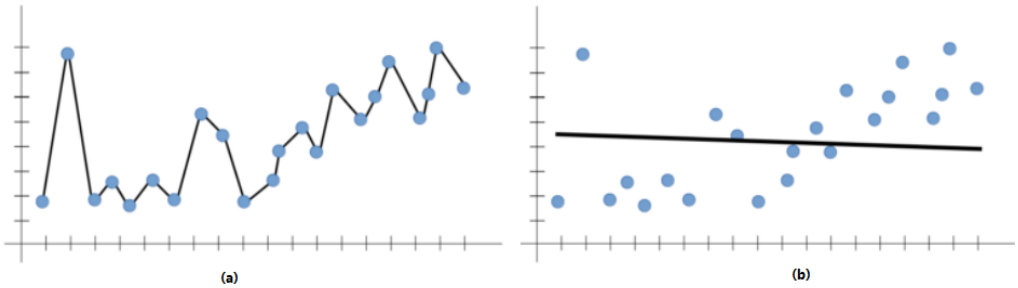


Figura 4. Muestra de un entrenamiento con overfitting (a), y otro con underfitting (b). [14]

1.4 Redes Neuronales en R

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico [10]. Hay numerosas librerías que permiten desarrollar modelos con redes neuronales en R cuyas diferencias radican en la complejidad de la red a modelar y la facilidad de uso por parte del usuario.

Una de estas librerías es KERAS, una librería que permite desarrollar de manera sencilla distintas estructuras de redes neuronales basadas en los avances en el campo de Deep learning [9]. Se apoya en otras librerías como TensorFlow o Theano. KERAS define las capas de la red como objetos con ciertas propiedades y construye la red como una secuencia de estos objetos. Además, permite incorporar avances de Deep learning como redes convolucionales o descenso del gradiente estocástico.

2. Datos

2.1 Descripción de los datos

El INE (Instituto Nacional de Estadística) es wl organismo encargado de la coordinación general de los servicios estadísticos de la Administración General del Estado y de la vigilancia, control y supervisión de los procedimientos técnicos de los mismos. Entre los campos en los que trabaja destacan la demografía, economía y sociedad. Este organismo es el encargado de la elaboración de los censos y padrones municipales.

Los censos de población y viviendas son recuentos exhaustivos de la población de un país que se realizan cada 10 años y que permiten conocer las características sociales y demográficas de las personas (edad, estudios…) y constituyen la operación de mayor rango dentro de la actividad estadística oficial. [11].

A partir de la información de los censos se pueden obtener conclusiones sobre el número de habitantes y su distribución, el modo de estructura de los hogares, el número de personas trabajando, estudiando, las características de las viviendas, locales y oficinas, etc. Lo que permite desarrollar políticas demográficas y evaluar los resultados de estas, así como asignar los recursos económicos por el territorio.

Los datos que se utilizan para este problema son proporcionados en diferentes ficheros con variables de los censos de 2001 y 2011 y variables territoriales. Estas últimas se obtienen por medio de distintos análisis de sistemas de información geográficos, SIG.

Este problema se basa en el análisis de los valores proporcionados por estos censos elaborados por el INE en el año 2001 y 2011. Para 2001 se tienen 5 ficheros diferentes, a continuación, se explicará brevemente el contenido de cada fichero:

- **Población vinculada y sus componentes**: en este fichero las variables hacían referencia a la población residente y a la población vinculada no residente clasificada según su vinculación con el municipio. Se denomina población vinculada como el conjunto de personas censables, es decir, con residencia habitual en España, que tienen algún tipo de vinculación habitual con el municipio ya sea porque trabajan allí, estudian allí…
- **Población según sexo y edad**: este fichero contiene información sobre el número de personas de un municipio en grupos quinquenales, así como el porcentaje de población mayor de 65 años y el índice de envejecimiento.
- **Indicadores territoriales municipales**: divide las variables en siete grupos: tamaño, vivienda, estructura demográfica, tendencia demográfica, hogares, actividad económica y nivel de vida y estudios.
- **Población en viviendas familiares según equipamiento del hogar**: variables sobre el equipamiento de las viviendas como la calefacción, instalaciones de refrigeración o vehículos en dicho hogar.
- **Núcleos familiares según tipo de núcleo**: número de madre con hijos, padre con hijos…

En el caso del censo de 2011, al ser un censo que proporcionaba menor información debido a su metodología de producción, solo se va a considerar en el presente trabajo una variable que indique la población total de cada municipio que se usará para definir si el municipio se despobló. En este caso existe un registro de 8.116 municipios.

Las variables territoriales se encuentran en un solo fichero que contiene variables sobre la distancia a carreteras nacionales, autopistas, altitud, etc. Para este fichero se cuenta con un total de 8.198 registros.

2.2 Carga de datos en R

El primer paso ha sido cargar estos ficheros en R, para ello se utiliza la librería (readxlx) [7]. Readxlx proporciona una manera de obtener datos a partir de un excel en R devolviendo un data frame. Esta librería soporta formatos .xlsx que son en los que se encuentran los ficheros a utilizar. Otra librería que se utiliza para la carga y el procesamiento de los datos en R es tidyverse [8].

Primero se han cargado los datos utilizando la librería xlsx , saltándose las filas necesarias de los encabezados de los ficheros

El procedimiento de carga y curación de los datos ha sido similar para todos los ficheros.

2.3 Curación de datos

La curación de datos es una fase necesaria dentro del ciclo de vida de los datos que garantizan la calidad del dato que posteriormente se utilizara en las técnicas de aprendizaje. Dentro de la etapa de curación de los datos se pueden distinguir dos tipos de problemas: problemas “single-source” y “multiple-source”.

Problemas “single-source” que se encuentran en los ficheros:

- **Errores ortográficos:** hay errores dentro del nombre del municipio de caracteres que no se leen correctamente al cargarse o que tienen discrepancia de idiomas.

- **Valores perdidos**: en algunos campos se encuentran valores vacíos.

- **Valores embebidos:** varios valores en un solo campo. Esto ocurre en algunos ficheros donde el código INE se encuentra junto con el nombre del municipio.

Problemas “multiple-source”:

- **Datos solapados:** dentro de los diferentes ficheros se encuentra varias veces el campo repetido de población total con distintos nombres.

- **Conflictos estructurales:** hay discrepancia en la forma de almacenar el código INE y el municipio. En algunos casos se encuentran en dos columnas separadas y en otros en una misma columna. Además, el nombre del municipio no sigue los mismos criterios en los diferentes ficheros.

Otro problema derivado de la utilización de datos provenientes de distintas fuentes es el problema de la UEM, Unidad Espacial Modificable, por cambios en las unidades administrativas de referencia en cada año. Este problema se observa en el número de municipios que se encuentra en cada censo, variando para 2001 y 2011.

Una vez analizados los datos para determinar la tipología de los que se ha abordado la etapa de curación y limpieza para garantizar su calidad.

El primer problema encontrado es que no todos los ficheros seguían el mismo convenio de nombres. Discrepancias en el uso de mayúsculas, la utilización de la "ñ", nombres en los distintos dialectos y orden del nombre hacen imposible juntar las tablas por este campo. Por ello se ha decido separar esta columna en dos, por una parte, el Código INE y por otra el nombre del municipio.

En algunos de estos ficheros a parte de los valores de los municipios también se encontraban registros de las provincias, estos valores se han eliminado ya que el estudio se realiza a nivel municipal.

Para todos los ficheros la variable con el código INE se ha llamado COD_INE para seguir un convenio. Este valor se trata de un código oficial establecido por el INE, con una longitud de 5 dígitos: los dos primeros correspondientes a la provincia y los restantes al municipio. Por último, se han guardado los distintos ficheros en formato csv.

Una vez todos los ficheros se encontraban en formato csv y siguiendo un criterio en cuanto al nombre de sus columnas, columna Código INE y Nombre del Municipio separadas, se juntan los ficheros del censo de 2001 a partir del Código INE, que se va a utilizar como Primary Key de las distintas tablas, ya que es un identificador único para cada municipio.

En algunas variables de 2001 de los distintos ficheros se repetía el valor de la población total por lo que se han eliminado todas excepto la variable que contenía esta información en el primer fichero. También se ha borrado el nombre del municipio para evitar conflictos.

A continuación, se cargan las variables territoriales, que han sido pasadas a formato csv siguiendo el procedimiento anterior. Existen dos valores iguales a 0 en la variable AREA KM2, al ser solo dos se ha buscado el valor del área de estos dos municipios (46117 y 46152) siendo en ambos casos el área igual a 0'4 km2, por lo que hemos sustituido el 0 por este valor.

Para el censo de 2011 consideramos solo el Código INE y la población total. Después se unen los datos de 2001, de 2011 y las variables territoriales en un mismo dataset cruzando las tablas por el COD_INE. Al existir en 2001 menos municipios registrados se cruzan las tablas cogiendo solo aquellos municipios que están registrados en todos los ficheros para tener la información completa y que no aparezcan nuevos campos vacíos. A este tipo de cruce de tabla se le conoce como "inner join".

Para determinar si un municipio se ha despoblado o no se utiliza la variable de densidad de población. Hay estudios que señalan un valor crítico de 8 habitantes por kilómetro

cuadrado pero la Comisión Europea establece un umbral crítico de 12’5 habitantes por kilómetro cuadrado, este último valor es el que se va a utilizar.

A partir de la variable de AREAKM2 de un municipio y las variables de población total de 2001 y 2011 se va a obtener la densidad poblacional en ambos años.

Por último, se han eliminado aquellos municipios con valores por debajo del umbral de la despoblación en 2001 ya que no aportan información debido a que es muy poco probable que un municipio ya despoblado sufra un aumento de población que le haga superar este umbral. Tras eliminar estos municipios se cuenta con un total de 4.416.

Una vez se tiene el dataset completo con todas las variables y todos los municipios se va a realizar la curación de vacíos. A continuación, se muestran las variables con valores nulos y el criterio que se ha seguido para sustituir o eliminar este valor:

- **IND_ENVEJ (Índice de envejecimiento:** Es la relación de población a partir de 65 con la población menor de 15 años. Por lo que los valores nulos corresponden a aquellos municipios sin población en un rango de 0 a 14 años. Para evitar estos valores se ha eliminado esta columna y se ha sustituido por el tanto por ciento de población menor de 14 años, ya que ya existía una columna con el tanto por ciento de mayores de 65.

- **Indicador de juventud:** solo aparece un municipio con este campo vacío por lo que se ha eliminado del dataset.

- **Indicador de reemplazamiento:** existe 20 municipios con este valor nulo, pero al ser un campo que no se puede calcular se eliminan dichos municipios.

- **Crecimiento de la población:** es la tasa de crecimiento medio interanual en %. Este indicador tiene una utilidad en el estudio de la despoblación, pero al no conocer detalladamente como se obtiene por lo que se eliminan los municipios con este valor vacío.

- **Incremento del número de edificios:** se eliminan los municipios con este campo vacío ya que no es posible calcular este valor.

- **Indicador de emancipación:** Es el porcentaje de personas de 30 a 34 años que siguen perteneciendo al núcleo de sus padres. Este indicador actualmente tiene más relevancia debido al retraso en la edad de emancipación sin embargo es un valor que no se puede calcular para los municipios en los que no aparece el valor. Por lo tanto, se eliminan los municipios con este campo vacío.

- **Media hijos por núcleo:** tan solo el municipio de Revilla Vallejera tiene este campo vacío. Analizando el municipio se ha encontrado que no ha habido ningún nacimiento en periodo intercensal, además de ser un municipio con poca población, por lo que tiene sentido sustituir el valor de media de hijos por núcleo por 0.

- **Parejas de hecho**: son cuatro municipios con poca población, el mayor con 128 habitantes, el resto no superan los 20. Se va a sustituir por 0 este valor ya que es un campo que no se puede calcular.

- **Estudios pre-obligatorios y Estudios post-obligatorios:** Una primera idea para reemplazar esta variable fue, para los que uno de los datos estaba informado intentar calcular el otro. Sin embargo, el tanto por ciento de estudios post y pre obligatorios no se ajusta al 100% ya que el denominador de ambos es distinto. Los estudios post-obligatorios son el porcentaje de personas entre 16 y 25 años que están cursando estudios de Bachillerato, COU, Escuela Oficial de Idiomas, FP, diplomatura, postgrado o equivalentes, respecto a la población de ese grupo de edad residente en el territorio.En el caso de los pre-obligatorios es el porcentaje de niños de 0 a 3 años que están escolarizados respecto al total de niños en ese rango de edad en ese territorio. Para estas dos variables se sigue el mismo procedimiento. Como no era posible determinar su valor de otra forma se ha hecho la media del valor para el resto de los municipios con poblaciones similares.

3. Análisis preliminar de las variables.

Para tener una primera aproximación sobre cómo se distribuyen los datos y si hay dependencias visibles entre las variables y la población de un municipio, se van a representar las variables para conocer sus distribuciones.

Para ello, a partir de la definición de umbral de despoblación, 12.5 hab/km2, se van a clasificar los municipios en poblados o despoblados para posteriormente representar sus variables para los dos conjuntos de datos y ver si la distribución de dichas variables varía en función de los dos subconjuntos que se han definido.

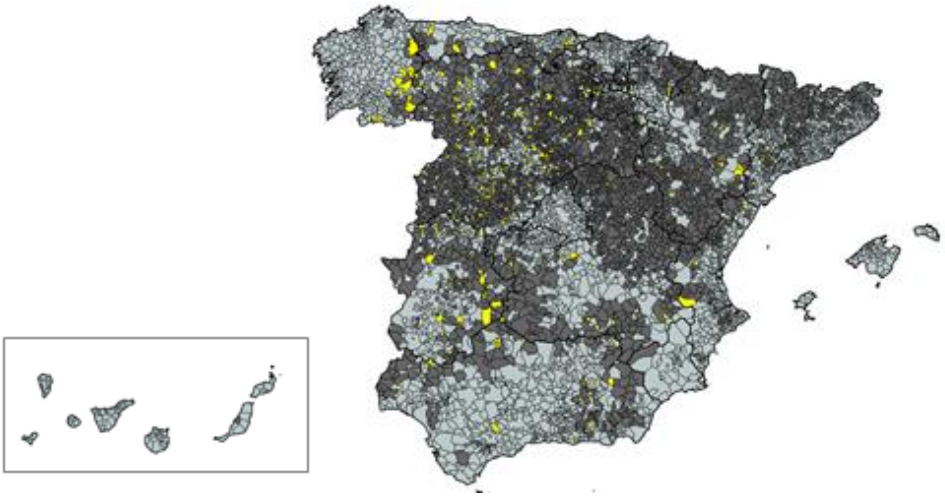


Figura 5 Mapa de España en el que se encuentran en gris oscuro los municipios por debajo del umbral crítico de densidad de población en 2001 y en amarillo en 2011.[16]

Con los municipios clasificados como poblados o despoblados en 2001 y 2011 se ha representado su distribución espacialmente dentro del territorio nacional. La **Figura 4**, muestra en gris oscuro los municipios despoblados en 2001 según el criterio que se ha determinado de densidad de población y en amarillo los despoblados en 2011. Se observa que siguen un patrón espacial coherente concentrándose en zonas del interior y de montaña y distribuyéndose los despoblados en 2011 en colindancia con los del periodo anterior.

Posteriormente se lleva a cabo la representación de cada variable distinguiendo entre los poblados y despoblados en 2011.

3.1 Variables territoriales

Se han encontrado valores extremadamente elevados para algunas variables por lo que se ha buscado en qué municipios ocurría esta circunstancia para ver si el valor tenía sentido o no. Las variables para las que se han obtenido estos valores son:

- **DISTAUTOPAUTOV_M:** distancia en metros del municipio a una autopista o autovía. En este caso el máximo valor encontrado, 162 km, pertenece a Melilla. El resto de los municipios con valores también superiores a los 100 km pertenecen a municipios de las Islas Canarias, principalmente a municipios de la

isla de El Hierro. Estos valores son debidos a que ni en El Hierro ni en Melilla hay autopistas por lo tanto el valor que se está cogiendo en el caso de El Hierro es el de la autopista más cercana en otra isla, y en el caso de Melilla el de una autopista de la península.

- **DISTCARRNAC_M:** distancia en metros a una carretera nacional. Los máximos valores encontrados están entorno a los 1000 km, y todos los municipios pertenecen a las Islas Canarias. Esto es debido a que en Canarias no hay carreteras nacionales, por tanto, la distancia que se está dando es la que hay de una autopista de la península.

- **DISTESTACFERROC_M:** distancia en metros a una estación de ferrocarril. Al igual que en el caso de la distancia a una autopista/autovía, los municipios para los que se encuentran valores muy elevados pertenecen a Melilla o algún municipio de las Islas Canarias.

La distancia a estaciones, carreteras nacionales o a una estación de ferrocarril en los casos mencionados anteriormente no son datos que aporte una información real, ya que en ninguno de los casos son puntos accesibles desde los municipios por su localización especial. Por ejemplo, la distancia a una estación de ferrocarril en una isla en la que no hay línea de ferrocarril no va a afectar al fenómeno de despoblación en la realidad, pero un valor tan elevado sí que puede afectar en el modelo de aprendizaje automático.

El criterio que se va a utilizar para tratar estos valores es sustituirlos por el valor máximo encontrado para el resto de los municipios. Es decir, para los municipios fuera de la península para los que los valores eran anómalos se ha sustituido por el máximo valor para un municipio dentro de la península.

3.2 Variables del censo 2001

En las variables que dependen de la población, el número de viviendas o de locales totales existen valores muy grandes en comparación con el resto que hacía que no sea clara su interpretación en los histogramas. Se ha buscado a que municipios correspondían estos valores y en todos los casos pertenecían a grandes núcleos urbanos del país, los cinco municipios con el valor más elevado: Madrid, Valencia, Sevilla, Zaragoza, Málaga. Para tratar estos datos, se ha calculado el porcentaje en función del total.

Por ejemplo, en el caso de la variable 'Sólo.reside', que hace referencia al número de personas que solo residen en un municipio, el valor máximo, que pertenecía al municipio de Madrid, era de 1669594 frente a la media que es de 4823

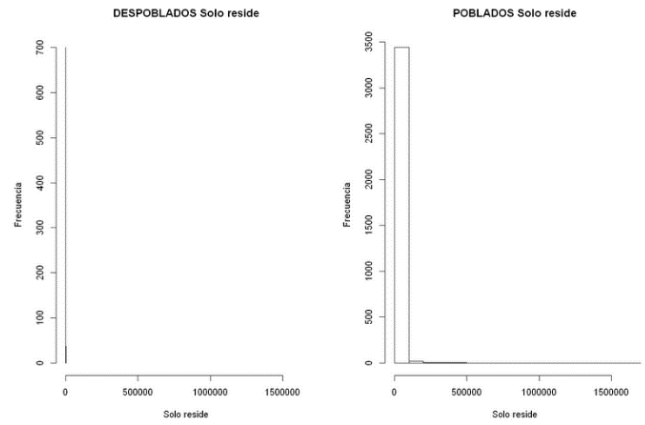


Figura 6. Histograma de la distribución de la variable "Solo reside" para municipios despoblados y sin despoblar en 2011.

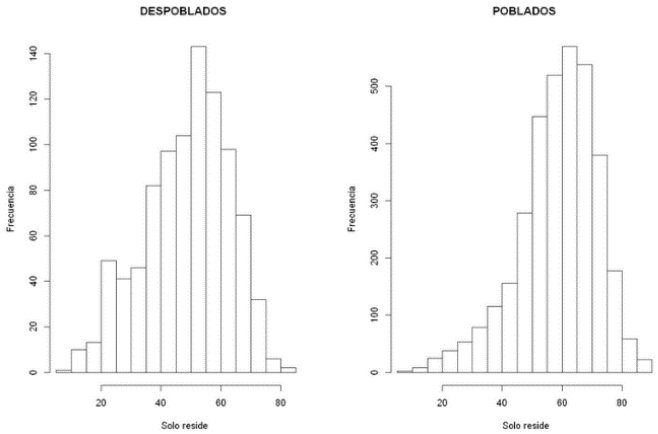


Figura 7. Histograma de la distribución de la variable "Solo reside" para municipios despoblados y sin despoblar en 2011 después de haber pasado los datos al % de población.

En el caso de las variables total viviendas, total locales y total edificios los máximos valores pertenecían también a los grandes núcleos urbanos. En este caso se ha tomado el logaritmo en base 10. Un caso especial ha sido la variable "Total Locales", ya que tenía valores iguales a 0 por lo que se ha tomado *log(x+1).*

3.4 Poder discriminante de cada variable

Los distintos histogramas obtenidos para la distribución de las variables en función de si se trata de municipios poblados o despoblados en 2011 ya proporciona información relevante sobre la importancia de dicha variable en la despoblación. Algunos histogramas que destacan se van a comentar a continuación:

-Variables territoriales:

- Altitud: altitud a la que se encuentra un municipio en metros. Para los despoblados existe una alta concentración para 700-1000 metros, bajas frecuencias en altitudes más favorables en despoblación. Se observa un alto

contraste con los poblados, que tienen altas frecuencias en el intervalo 0-500m y disminuye a medida que aumenta la altitud.

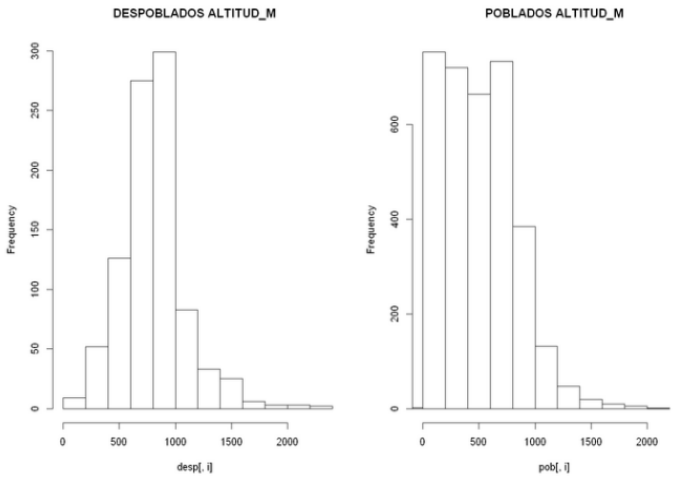


Figura 8. Histograma de la distribución de la variable "Altitud m" para municipios despoblados y sin despoblar en 2011.

- Distancia a núcleos de más de 10000 habitantes: la distancia está expresada en metros. Se observa también una diferencia en las distribuciones de los municipios poblados y los despoblados en relación a la distancia a estos núcleos que por su entidad o tamaño demográfico concentran una determinada cantidad de equipamientos y servicios a la población. Los despoblados apenas tienen casos en el primer intervalo de distancia, frente a los poblados que es el intervalo para el que tienen un mayor número de municipios.

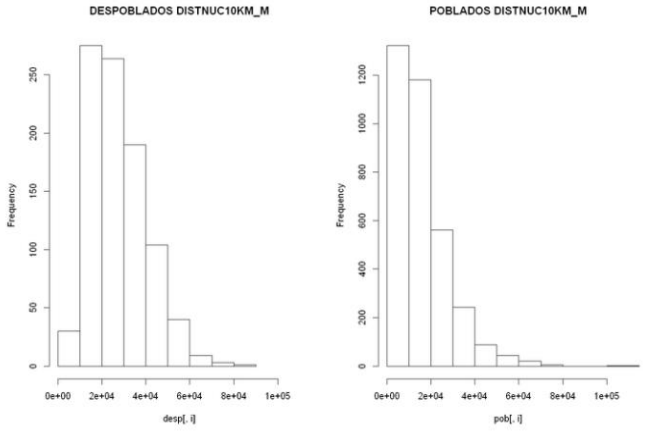


Figura 9. Histograma de la distribución de la variable "Distancia nucleo 10K_M" para municipios despoblados y sin despoblar en 2011.

- Distancias autovías: mucho más concentrada su distribución en los poblados elación con la cercanía. No es muy marcada la diferencia por la amplia red de tramos que tenemos.

- Distancia carreteras nacionales: Los despoblados no están muy alejados de las carreteras nacionales, lógico si se tiene en cuenta el sistema de asentamiento de nuestro país.
- Estaciones de ferrocarril: similar a las carreteras, influencia muy desigual según las partes del territorio consideradas.

- Variables estructurales (edades): manifiestan claramente la importancia de conexión entre despoblación y proceso de envejecimiento por la base y la cúspide de la pirámide.

- % Población mayor. En despoblados bastante superior (30-40%) frente a poblados (15-25%). En los poblados hay que tener en cuenta que arrastramos el envejecimiento propio de los municipios urbanos centrales, pero a pesar de esto el salto de valor es interesante.

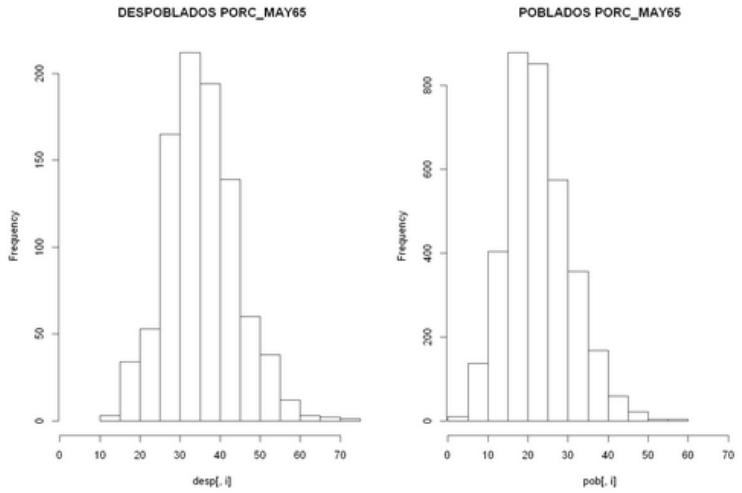


Figura 10. Histograma de la distribución de la variable "PORC_MAY65" para municipios despoblados y sin despoblar en 2011.

- % Población joven. En despoblados presenta niveles bajos (8-10%) frente a poblados que rondan 12-17%. Avala el envejecimiento por la base de la pirámide.

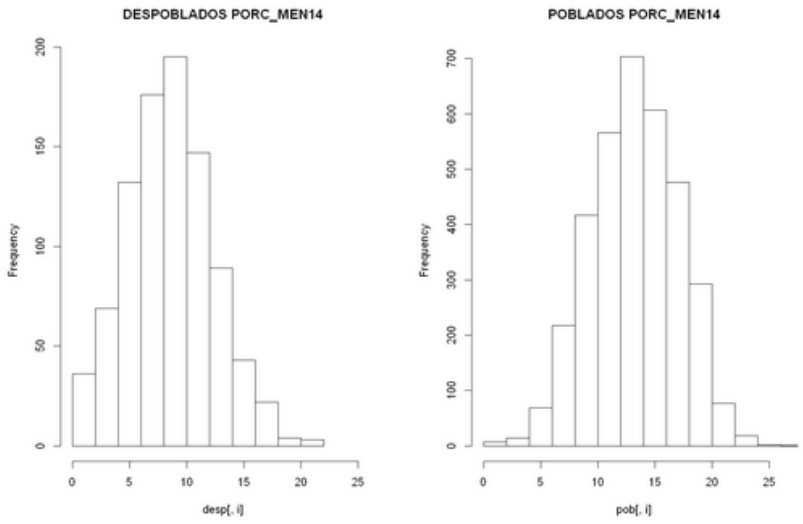


Figura 11. Histograma de la distribución de la variable "PORC_MEN14" para municipios despoblados y sin despoblar en 2011 después de haber pasado los datos al % de población.

- Edad media: expresivo con un contraste de los modales de los dos colectivos de municipios de unos 10 años de edad media.

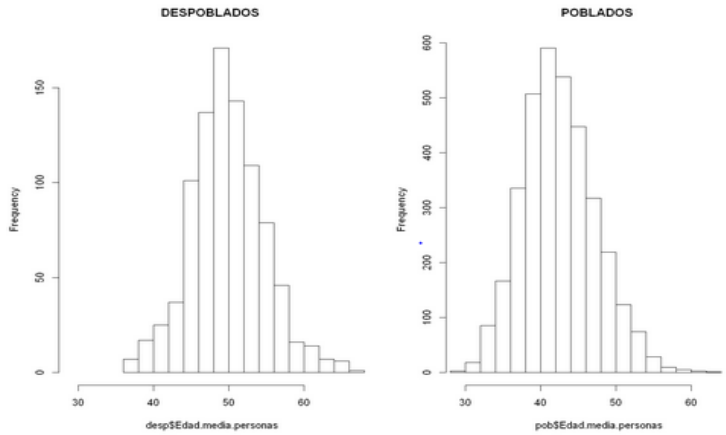


Figura 12. Histograma de la distribución de la variable "Edad media" para municipios despoblados y sin despoblar en 2011.

A partir de los resultados obtenidos tras la elaboración de los histogramas cabe destacar la gran diferencia en la distribución para municipios poblados y despoblados de los indicadores de edad de la población (% mayores de 65, % menores de 14), así como la distribución de la variable altitud, acumulándose para los municipios poblados la mayor parte en los primeros rangos (valores inferiores a 1.000 metros de altura).

4. Diseño de una Red Neuronal

El objetivo de este estudio es poder conocer la despoblación de un municipio en un periodo de 10 años a partir de las variables que se recogen en los censos y las variables territoriales, que se suponen como constantes a lo largo del tiempo ya que son variables como la altitud, la pendiente o la distancia a grandes núcleos urbanos. Además, se quiere determinar qué factores influyen más en la predicción de la población.

Las predicciones se van a realizar sobre la población de los municipios en 2011 y como variables predictoras se va a usar la información del censo de 2001, al ser este más completo, y las variables territoriales.

La variable por la que se va a determinar si un municipio se ha despoblado o no es la densidad de población, por debajo de una densidad de 12.5 hab/km2 se considerará como despoblado. La predicción se va a abordar como un problema de clasificación, por lo tanto, la variable de densidad de población de 2011 se va a pasar a una variable binaria tomando el valor 0 cuando el valor esté por debajo de 12.5 y 1 cuando sea igual o mayor a este valor.

Debido a que la predicción se va a realizar a partir de un gran número de variables se ha determinado que el procedimiento más adecuado es aplicar el modelo de redes neuronales. El modelo de red neuronal se va a realizar a partir de la librería KERAS [10] en el entorno R [11].

3.1 Construcción de la red neuronal

A continuación, se va a detallar el proceso para la construcción del modelo que se usará posteriormente para realizar la predicción sobre la despoblación.

El primer paso es proporcionar la estructura adecuada a los datos, que previamente ya se han sometido a una etapa de curación explicada en el Capítulo 2, ya que la red neuronal no es capaz de procesar campos vacíos.

Para el desarrollo de la red neurona se tienen 4382 municipios y 109 variables: 108 variables predictoras y la variable objetivo (despoblación en 2011). El dataset que contiene las variables predictoras se van a denominar, ***x***, y el dataset con la variable objetivo, ***y.***

Es importante escalar las variables para evitar problemas de estabilidad y mejorar el rendimiento del entrenamiento. Los datos de las variables predictoras se van a escalar a partir de la función “scale” de KERAS. Esta función centra y escala las columnas de una matriz. Cada columna de la matriz se centra restándola en valor de su media. El escalado de las variables se realiza dividiendo la columna, previamente centrada, por su desviación estándar. El dataset ***x*** tras el proceso de escalado de los datos será lo que constituya la capa de entrada de la red neuronal. En el caso de la variable ***y*** al ser un problema de clasificación, solo toma valores de 0 ó 1, no es necesario el escalado.

Una vez se ha preparado el dataset de las variables predictoras hay que proporcionar la estructura necesaria al dataset ***y***. La variable que se quiere predecir debe tener la misma estructura que la capa de salida de la red neuronal.

Se ha determinado que la capa de salida óptima sea aquella que tenga el mismo número de neuronas como de clases a identificar. Al ser un problema de clasificación binaria la salida va a estar constituida por dos neuronas. La primera neurona determina la probabilidad de que el caso que se esté prediciendo sea 0, es decir despoblado, y la segunda de que sea 1, poblado.

Para pasar la variable objetivo ***y*** a esta estructura se convierte en una matriz de dos dimensiones donde la primera dimensión es la observación y la segunda la neurona oculta.

Por último, ***x*** e ***y*** se van a dividir en dos subconjuntos, el 80% se va a utilizar como muestra de entrenamiento y el 20% restante como test. Los municipios que van a pertenecer a cada muestra se escogen de manera aleatoria ya que al estar ordenados en los censos por provincias podían estar sesgados.

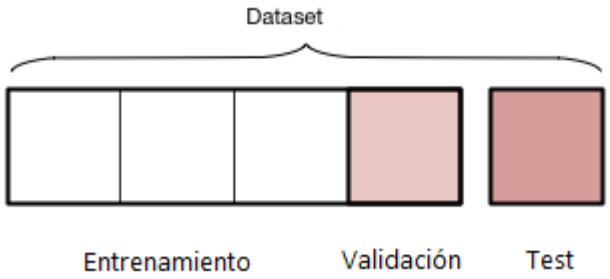


Figura 13. Esquema sobre los subconjuntos en los que se divide el dataset inicial.

Una vez se tienen los datos con la estructura deseada el siguiente paso es la construcción del modelo de redes neuronales.

KERAS permite desarrollar el modelo de forma secuencial o de manera personalizada. Se ha determinado desarrollarlo de manera secuencial, la red va a ser una secuencia lineal de capas. Primero se define el tipo de capa, el número de neuronas ocultas que va a tener cada capa y el tipo de activación.

Como parte del entrenamiento del modelo hay que determinar el error del modelo en cada iteración para poder ajustar los pesos en cada iteración y mejorar la salida del modelo. Para ello se introduce una función de coste que se usará para medir el error del modelo. En este caso la función que se va a emplear es el error cuadrático medio (“mean square error”), que calcula la media de la diferencia cuadrática entre lo predicho y el valor real. El resultado perfecto sería obtener que la función llegará a 0 por lo que al entrenar el modelo se va a intentar minimizar esta función.

Debido a que el cálculo de la derivada parcial de la función de coste respecto a cada uno de los pesos de la red para cada iteración tendría un elevado coste computacional se introduce un optimizador, en este caso se va a empelar el optimizador SGD (Stochastic Gradient Descent). SGD limita el cálculo de la derivada a solo una observación por iteración.

El gran número de parámetros que una red neuronal debe ajustar ocasiona que estos modelos tengan tendencia a sobreajustar cuando el número de datos en el entrenamiento no es suficiente. Además, al entrenar sobre la muestra de entrenamiento la red aprende

como minimizar el error en esta, pero puede que el error no se minimice en la misma dirección en la muestra de test y por tanto que la red no tenga poder de generalización.

Para solucionarlo se utiliza una muestra de validación que controle la divergencia de errores. En este caso un 10 % de la muestra de entrenamiento se usará como validación, esta fracción será apartada del modelo para que no se entrene con estos datos y evaluará la función de coste y el accuracy para dicha muestra en cada iteración. Este procedimiento sirve para verificar que el entrenamiento está siendo adecuado y que la red neuronal puede generalizar y realizar predicciones correctas sobre datos con los que no está aprendiendo el modelo.

Se utilizan distintas técnicas para poder determinar la validez del modelo.

Una forma de determinar la validez del modelo es representar la curva de entrenamiento tanto de la muestra de entrenamiento como la de validación para observar si el modelo tiene overfitting, underfitting y cuantas épocas son adecuadas.

En la curva de entrenamiento se muestra la evolución tanto del accuracy, que calcula la frecuencia con la que la predicción es igual que la etiqueta, como de la loss-function para la muestra de entrenamiento y de validación. Se considera el mejor modelo aquel que obtiene mejores resultados para la loss-function en la muestra de validación.

También se ha obtenido la curva ROC, utilizando las predicciones que el modelo hace a partir de la muestra de test, que es una representación de la razón de verdaderos positivos frente a falsos positivos. A parte de la información que se obtiene gráficamente sobre la validez del modelo con la curva ROC también se calcula el área bajo la curva, AUC, resultados más elevados de este valor indican una mejor clasificación.

Una vez determinada la función de coste y el optimizador que se van a emplear se van a realizar pruebas con distintas estructuras y parámetros para determinar el modelo que mejor resultados proporcione. Los parámetros a ajustar son el número de capas y neuronas ocultas, el número de épocas y el valor de learning rate.

3.2 Ajuste de parámetros

Inicialmente se desarrolló un modelo de red neuronal con 3 capas ocultas con 100 neuronas en las dos primeras y 2 en la capa de salida. Al ser un modelo complejo y no tener un gran número de datos para entrenarlo no se obtuvieron resultados adecuados. En la curva de entrenamiento se observa como la función de coste para la muestra de entrenamiento no ha llegado a estabilizarse y sigue aprendiendo.

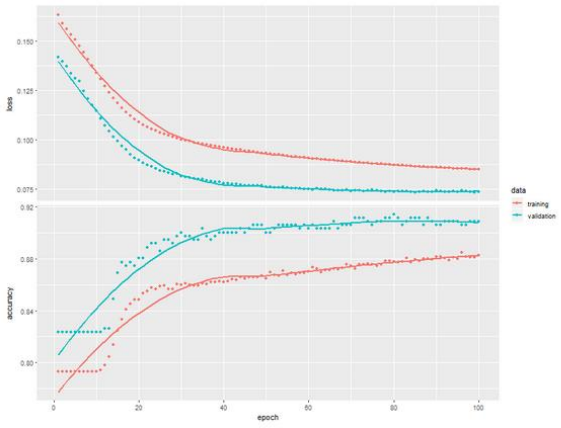


Figura 14. Curva de entrenamiento para el modelo con todas las variables y todos los municipios. En naranja se observa los valores obtenidos para la muestra de entrenamiento y en azul para la muestra de validación.

Además, analizando la predicción del modelo se ha observado que no logra predecir municipios como despoblados. Aunque la función de coste y el accuracy no proporcionen resultados extremadamente negativos esto es debido a que existe un número bastante superior de municipios poblados que despoblados por lo que el modelo solo logra predecir todos como poblados para minimizar el error. Por lo tanto, se determina que esta estructura no es válida.

Se ha intentado reducir la complejidad del modelo construyendo una red neuronal con 2 capas densamente conectadas, una de ellas es una capa oculta con 20 neuronas y la otra la capa de salida con 2 neuronas. La función de activación de ambas capas va a ser la función sigmoide.

```
Model
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 20)                2180
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 42
=================================================================
Total params: 2,222
Trainable params: 2,222
Non-trainable params: 0
_________________________________________________________________
```

Figura 15. Esquema de la estructura del modelo de redes neuronales.

Este modelo con una estructura más simple obtiene resultados mejores que el anterior. Sin embargo, la diferencia entre la función de coste para la muestra de validación y la de entrenamiento era bastante elevada cuando se comenzó a entrenar el modelo con 100 épocas, siendo superior el valor de la función de coste para la muestra de entrenamiento. Esto indicaba que el modelo no conseguía reducir el error y por tanto había underfitting.

Para intentar solucionar este problema se entrenó el modelo durante 250 épocas obteniendo así finalmente valores muy similares para ambas muestras. En el accuracy,

en cambio, sí que se observa que la muestra de entrenamiento da mejores resultados que la de validación.

5. Clasificación

4.1 Primer modelo – Todas las variables y todos los municipios.

Una vez determinada la estructura de la red y el valor de los parámetros que se van a utilizar se comienza con el estudio de los resultados que obtiene el modelo y la validez de ellos. El problema se ha comenzado a abordar con todas las variables del censo de 2001 y las variables territoriales, un total de 108 variables, como predictoras y con la población de 2011 como variable objetivo. El número total de municipios que se tienen en el dataset son 4.382.

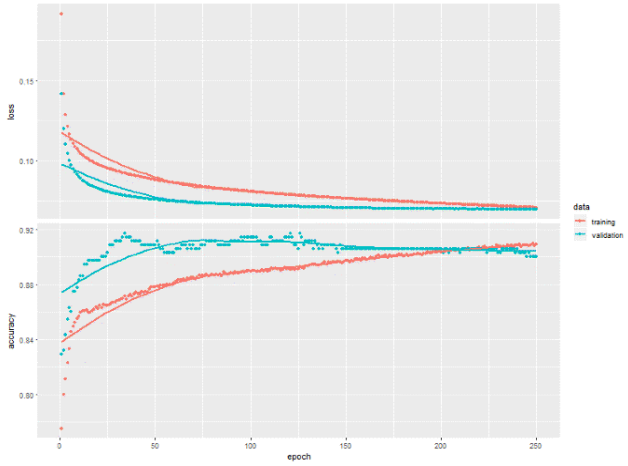


Figura 16. Curva de entrenamiento para el modelo con todas las variables y todos los municipios. En naranja se observa los valores obtenidos para la muestra de entrenamiento y en azul para la muestra de validación.

En la **Figura 16** . se puede observar cómo durante el entrenamiento del modelo se minimiza la función de coste y aumenta el accuracy.

$\mathbf{Loss_t}$	$\mathbf{Loss_v}$	$\mathbf{Accuracy_t}$	$\mathbf{Accuracy_v}$
0’0752	0’0699	0’9036	0’9117

Tabla 1. Resultaos de loss y validation obtenidos para ambas muestras para el modelo con todas las variables, todos los municipios y un entrenamiento de 250 épocas.

A continuación, se muestra la curva ROC obtenida para este modelo, y el AUC calculado:

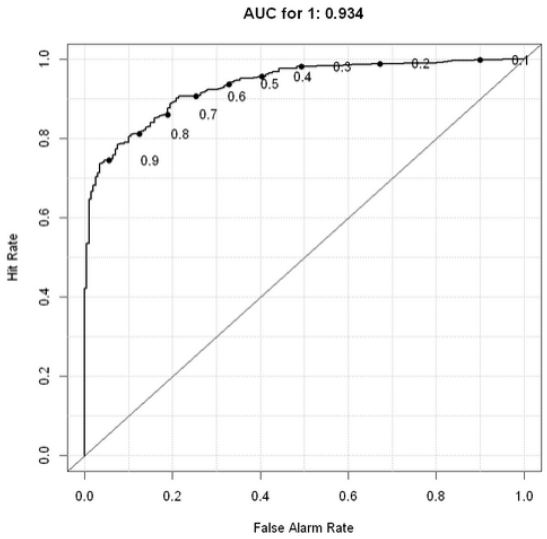


Figura 17. Curva ROC para el modelo de red neuronal con todas las variables y todos los municipios. En el eje **x** se encuentran el ratio de falsos positivos y en el eje **y** el ratio de valores predichos como positivos.

Con este modelo se ha conseguido obtener un valor relativamente bajo de función de coste tanto para la muestra de entrenamiento como para la de validación, además de un alto valor de accuracy. Los buenos resultados que proporciona el modelo también se pueden observar en la curva ROC, donde se puede observar que para el umbral de 0'5 más del 90% de los positivos se clasifican correctamente. También se obtiene un valor muy elevado de área bajo la curva, que está relacionado con la capacidad del modelo de clasificar correctamente un municipio.

Sin embargo, al modelo le cuesta más predecir aquellos municipios despoblados, para este umbral predice alrededor del 30% como falsos positivos, por lo que solo consigue clasificar correctamente el 70% de los despoblados.

Al haber municipios con poblaciones muy grandes es fácil determinar estos como poblados y el principal objetivo de este estudio es poder conocer que municipios con poblaciones más pequeñas se van a despoblar, ya que es muy poco probable que municipios con grandes poblaciones se desocupen en un periodo de 10 años.

4.2 Segundo modelo – Todas las variables y municipios < 50000 habitantes

Para que la red no se centre en aprender como clasificar casos en los que resulta obvio que no se van a despoblar en un periodo de 10 años se eliminan los municipios con poblaciones mayores a 50 mil habitantes, que es el valor que el Ministerio de Fomento utiliza para separar los principales núcleos urbanos. [12]

De los 4382 municipios que se tenían inicialmente se ha reducido a 4271 al eliminar aquellos con poblaciones mayores a 50.000, es decir, un 2'5% de los municipios pertenecían a grandes núcleos urbanos.

La red neuronal utilizada para entrenar este modelo tiene la misma estructura y los mismos valores para los parámetros que la anterior, sin embargo, el número de épocas se ha reducido a 100 ya que así se obtenían mejores resultados.

A continuación, en la **Figura 18** se muestra la curva de entrenamiento para este modelo.

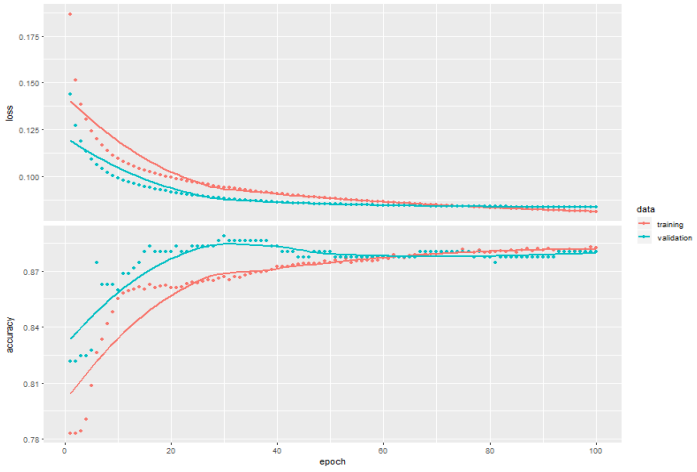


Figura 18 . Curva de entrenamiento para el modelo con todas las variables y los municipios con poblaciones menores a 50.000 habitantes. En naranja los valores obtenidos para la muestra de entrenamiento y en azul para la muestra de validación.

En la curva de aprendizaje para este modelo se puede observar cómo existe un punto donde los resultados se vuelven estables tanto para la muestra de entrenamiento como para la de validación y la diferencia entre ellas es pequeña. Esto indica que la curva de aprendizaje es correcta.

Para este modelo se han obtenido los siguientes resultados:

Losst	Loss v	Accuracy t	Accuracy v
0’0812	0’0838	0’8826	0’8801

Tabla 2. Valores obtenidos de función de coste y accuracy para la muestra de entrenamiento y de validación.

A continuación, se muestra la curva ROC obtenida para este modelo, y el AUC calculado:

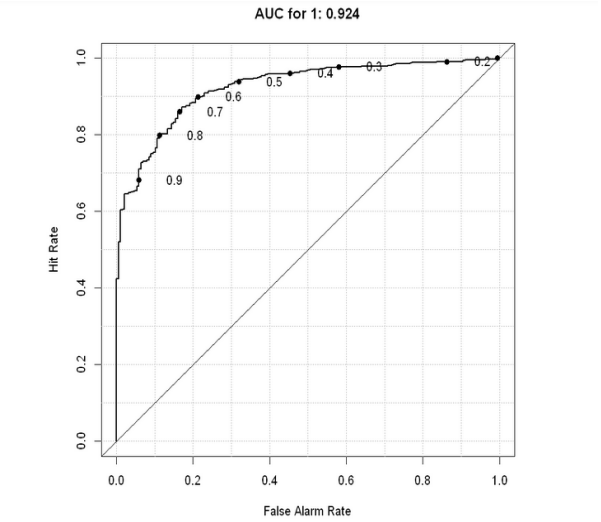


Figura 19 Curva ROC obtenida para el modelo con todas las variables y los municipios con poblaciones menores a 50000 habitantes.

Se ha obtenido un AUC de 0’924, es decir, existe un 92’4% de probabilidad de que la predicción de que un municipio este poblado sea cierta.

Para este modelo también se ha obtenido un resultado satisfactorio para la función de coste en ambas muestras, aunque es ligeramente superior al del modelo anterior. La curva ROC muestra que se puede lograr un área de trabajo con más del 95 % de aciertos y alrededor de 50% de falsos positivos. Además, en el otro extremo de la gráfica se observa que se puede obtener el 70% de aciertos sin apenas falsos positivos. Para un umbral de 0’5 se consigue predecir correctamente más del 90% de los municipios poblados con un 30 % de falsos positivo, es decir predice correctamente el 70 % de los despoblados.

4.3 **Reducción de variables.**

A veces un gran número de variables puede introducir mucho ruido en la red si estas no aportan información relevante por tanto se pretende reducir el número de variables para determinar si hay algunas que no aportan información y si al eliminarlas se obtienen mejores resultados.

Para ello se quiere obtener un modelo con aquellas variables que tienen una mayor influencia en la salida de la red. Esta dependencia se determina por medio de dos métodos

Correlación spearman:

La correlación de Spearman es una medida sobre la correlación entre dos variables aleatorias que oscila entre -1 y 1, en función de si las variables están relacionadas negativa o positivamente

A partir de la predicción obtenida por el modelo explicado anteriormente para todas las variables de 2001 y aquellos municipios con poblaciones menores de 50.000 se ha determinado la correlación de spearman de la predicción con cada una de las variables. Para obtener solo un dataset con aquellas variables que mayor dependencia tienen se ha decido quedarse solo con aquellas que tienen un valor de correlación mayor o igual que

0’5 en términos absolutos. El número de variables se reduce un 35% pasando de 108 a 38.

A continuación, en la **Figura 20** se muestra la curva de entrenamiento del nuevo modelo.

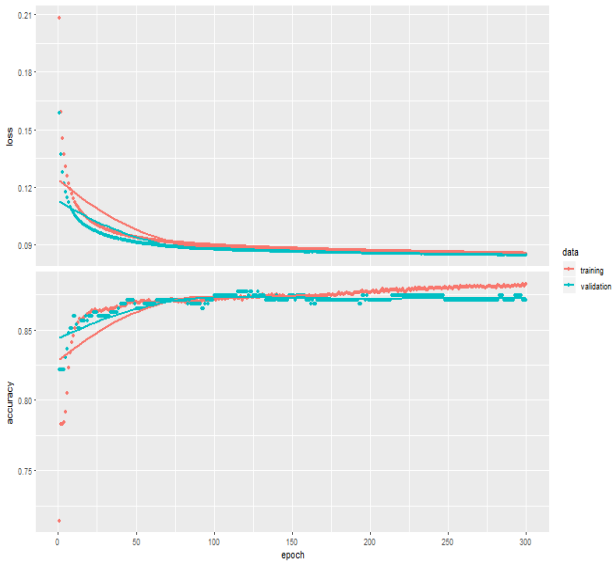


Figura 20. Curva de entrenamiento para el modelo para la reducción de variables y los municipios con poblaciones menores a 50.000 habitantes. En naranja los valores obtenidos para la muestra de entrenamiento y en azul para la muestra de validación

Para este modelo se han obtenido los siguientes resultados:

$\mathbf{Loss_t}$	$\mathbf{Loss_v}$	$\mathbf{Accuracy_t}$	$\mathbf{Accuracy_v}$
0’0844	0’0821	0’8803	0’8889

Tabla 3. Valores obtenidos de función de coste y accuracy para la muestra de entrenamiento y de validación.

Gráficamente:

El otro procedimiento que se ha utilizado para determinar qué variables están más relacionadas con la salida de la red ha sido representar la predicción en función de las diferentes variable.

Para la representación de la salida de la red en función de las distintas variables se han utilizado los gráficos boxplots que permiten observar la distribución de la muestra. En este caso se va a representar la distribución de la predicción de la red en función de distintos rangos de valor de la variable. La predicción toma valores entre 0 y 1, cuando la predicción es menor de 0.5 indica que el municipio está despoblado y mayor poblado. Las líneas gruesas que se observan en los gráficos (**Figura.21**) indican la mediana de la distribución, la parte inferior de la caja el primer cuartil y la superior el tercer cuartil.

Ejemplo de los profiles de las variables con dependencia con la salida.

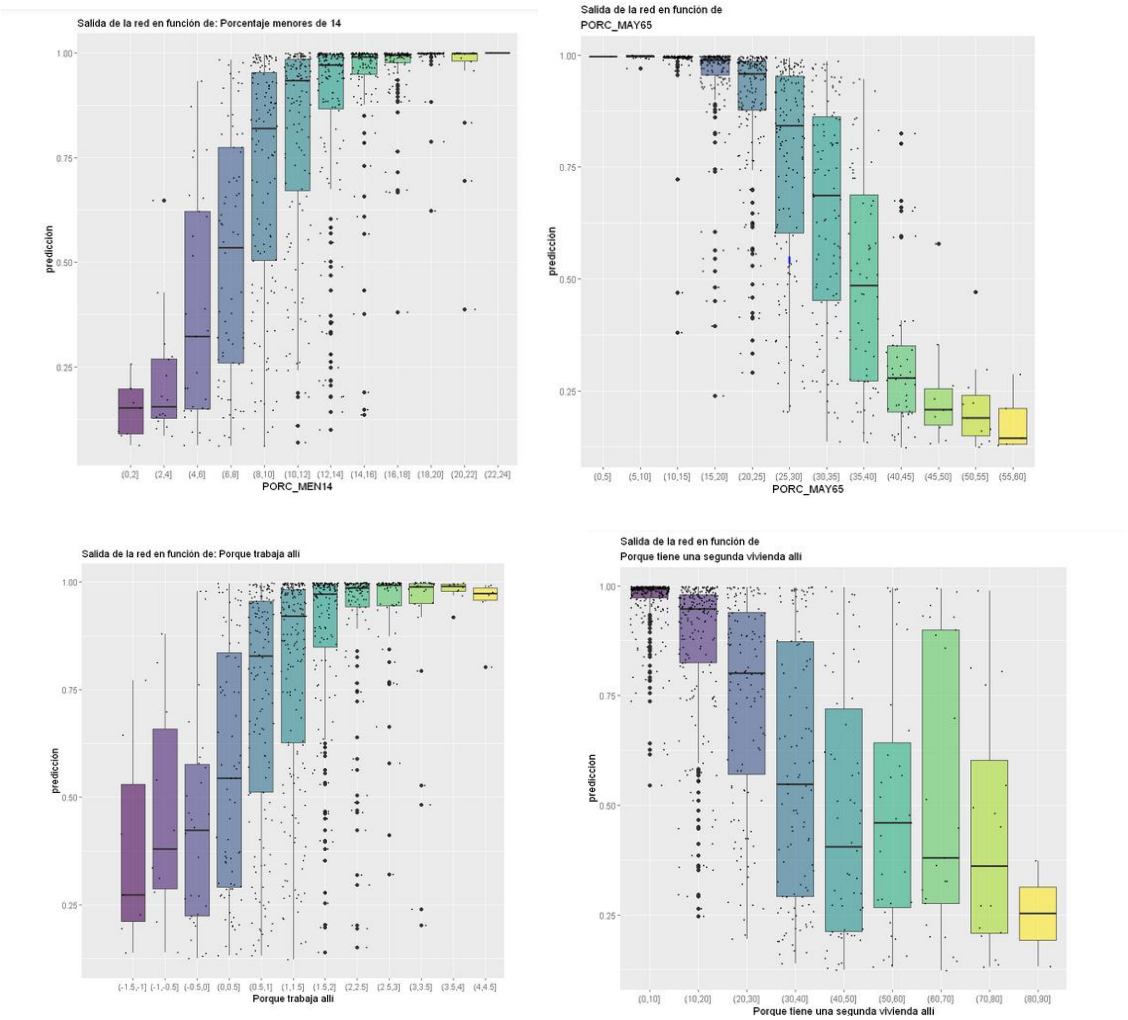


Figura 21 . Profiles obtenidos para cuatro variables distintas en las que se observa dependencia con la salida. En el eje x se representan los valores de la variable y en el eje y la predicción.

De las 108 variables que había inicialmente se han seleccionado 46.

A continuación, en la **Figura 22** se muestra la curva de entrenamiento para este modelo.

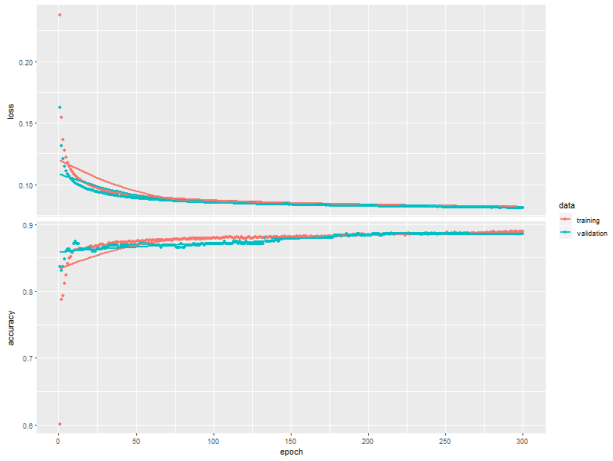


Figura 22 . Curva de entrenamiento para el modelo para la reducción de variables y los municipios con poblaciones menores a 50000 habitantes. En naranja los valores obtenidos para la muestra de entrenamiento y en azul para la muestra de validación

Para este modelo se han obtenido los siguientes resultados:

$Loss_t$	$Loss_v$	$Accuracy_t$	$Accuracy_v$
0’0814	0’0813	0’8894	0’8860

Tabla 4. Valores obtenidos de función de coste y accuracy para la muestra de entrenamiento y de validación.

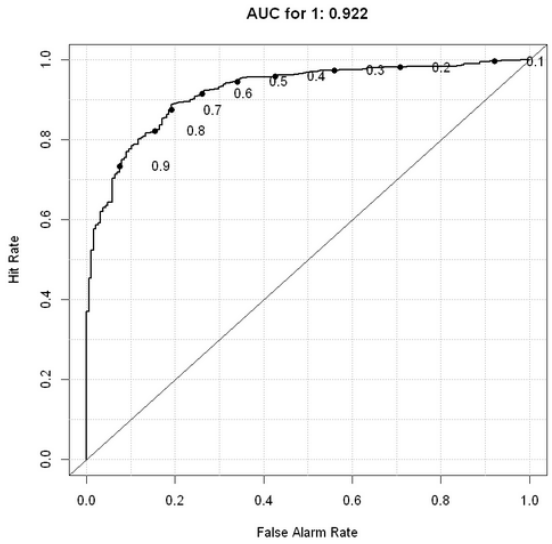


Figura 23. Curva ROC obtenida para el modelo con todas las variables y los municipios con poblaciones menores a 50.000 habitantes.

La función de coste obtenida para este modelo aun siendo bastante similar al modelo que utiliza todas las variables se ha conseguido reducir ligeramente para la muestra de validación. También la curva ROC ofrece buenos resultados en cuanto a la validez del

modelo, así como el valor del área bajo la curva, 92’2 %. En la Figura se observa que para un umbral de 0’5 predice correctamente más del 90 % de positivos con algo más del 30 % de falsos positivos.

4.4 Comparativa de los modelos.

En todos los modelos se obtienen resultados satisfactorios para la predicción de la población ya que se consigue clasificar los municipios con un alto valor de precisión. Aunque el que mejores resultados proporciona inicialmente es el que utiliza las 108 variables y todos los municipios hay que tener en cuenta que está utilizando núcleos con grandes poblaciones, superiores a 50.000 habitantes.

En el caso del resto de modelos hay que tener en cuenta que, aunque los resultados son ligeramente peores, se puede asumir que ese 2.5% de municipios que se han eliminado de este modelo los predeciría correctamente también, además obtiene mejores resultados para identificar los municipios despoblados.

Con la reducción de variables se han obtenido resultados igual de satisfactorios que para el caso anterior, utilizando la correlación de Spearman de las variables con la salida de la red, se han reducido las variables de 108 a 38. Analizando la distribución de la predicción en función de la variable, se han reducido las variables de 108 a 46, este método es el que mejor resultados ofrece tomando como indicador el valor de la función de coste para la muestra de validación.

Por tanto, el modelo determinado como aquel que proporciona un mejor valor predictivo de la despoblación es aquel constituido por 2 capas ocultas con 20 neuronas en la primera y 2 en la capa de salida, que utiliza como algoritmo de aprendizaje el descenso del gradiente estocástico y que se alimenta con un dataset de 46 variables, determinadas gráficamente como aquellas con mayor relación con la salida de la red, y con municipios con poblaciones inferiores a los 50.000 habitantes.

6. Análisis para la red neuronal seleccionada

Para el posterior análisis se va a utilizar el modelo que se ha determinado anteriormente como aquel que proporciona mejores resultados.

Este modelo se va a emplear para realizar un análisis más detallado de las variables y su dependencia.

6.1 Variables con más dependencia para esta red

Al ser la red seleccionada aquella que utiliza solo las variables con más relación con la salida de la red al representar de nuevo la predicción en función de cada variable todas ellas muestran dependencia. Sin embargo, siguen existiendo variables que muestran una mayor correlación.

Hay que tener en cuenta que se está estudiando la relación que tienen estas variables con la salida de la red y que no se puede identificar si esta relación indica la causa de la despoblación o es un efecto de ella.

A continuación, se va a analizar de forma más detallada esta correlación con la salida de la red neuronal para el modelo seleccionado, para ello se van a utilizar los gráficos "bloxplots" para la obtención de los profiles, igual que los gráficos que se utilizaron para determinar las variables con mayor dependencia. En algunos casos se ha pasado la variable a escala logarítmica para obtener una mejor visualización que proporcione más información. Con estos gráficos además se pretende poder determinar un punto de inflexión donde la muestra cambie su predicción, a este valor se le denominará como valor de "turn-on".

Se han seleccionado alguna de las gráficas más interesantes para mostrarse en este capítulo, el resto se encuentran en el **Anexo.**

-hab_km_2001: densidad de población en 2001. Campo calculado a partir de la población total y el área en km2 del municipio. Se observa que a partir de un valor de densidad en 2001 ningún municipio se predice como despoblado. La densidad de población se encuentra en escala logarítmica.

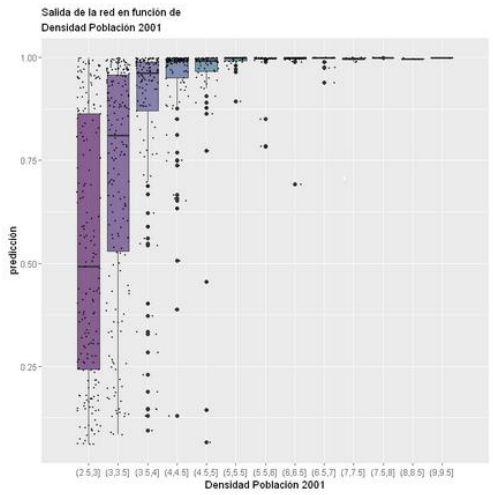


Figura 24. Distribución de la probabilidad de despoblación (siendo 1 no despoblado y 0 despoblado) frente a la densidad de población en 2001 en escala logarítmica.

-TOTAL...3(% población residente): porcentaje de población residente respecto del total de población de un municipio. A medida que aumenta disminuye la despoblación, por encima del 30% la mayor parte de los municipios se predicen como poblados, aunque siguen existiendo municipios despoblados.

-Porque tiene una segunda vivienda allí: forma parte de las variables de vinculación de población. Indica el porcentaje de población respecto al total que no reside en un municipio, pero está vinculada a él porque tiene allí una segunda vivienda. En el profile se puede observar que la probabilidad de despoblación de un municipio aumenta cuando el porcentaje de población vinculada por una segunda vivienda aumenta.
También se puede determinar un valor de "turn-on" a partir del cual la predicción de los municipios cambia, este sería alrededor del 40%.

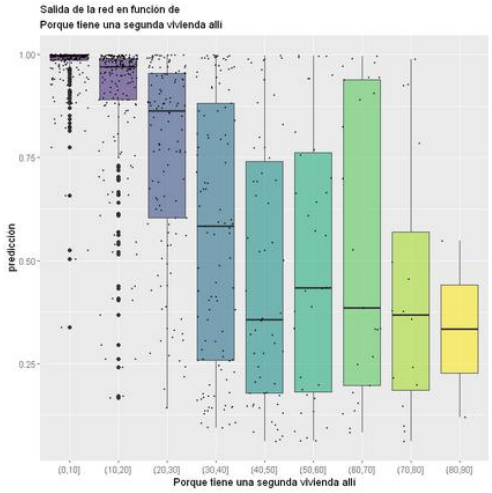


Figura 27. Distribución de la probabilidad de despoblación (siendo 1 no despoblado y 0 despoblado) frente al % de población vinculada porque tiene una segunda vivienda.

-Porque trabaja allí: al igual que el anterior forma parte de las variables de vinculación de la población. En este caso, es el tanto por ciento de población frente al total de población vinculada a un municipio que trabaja en el pero no reside. La representación se ha hecho en escala logarítmica ya que el rango de valores es muy amplio y la mayoría de los despoblados se acumulaban en la misma franja de valores por lo que no aportaba información. Al hacer la representación con la variable en escala logarítmica se puede observar como la probabilidad de despoblación disminuye a medida que el porcentaje de gente que está vinculada a un municipio por el trabajo aumenta.

- Solo.reside: % de población vinculada que reside en un municipio, ni trabaja ni estudia allí. En este caso observando el profile del valor de la predicción en función de la variable se observa como aumenta la probabilidad de despoblación a medida que aumenta el valor de la variable. Aunque se observa una dependencia para % altos de esta variable también existen un gran número de municipios despoblados por lo que no se puede determinar claramente un valor de "turn-on".

- **Total 8:** porcentaje de la población que está vinculada a un municipio pero no reside en él. A medida que aumenta el valor aumenta la probabilidad de despoblación de un municipio.

-Rangos de edad: De.0.a.4, De.5.a.9, De.10.a.14, De.15.a.19, De.20.a.24, De.25.a.29, De.30.a.34, De.35.a.39, De.40.a.44, De.50.a.54, De.45.a.49, De.55.a.59, De.70.a.74, De.80.a.84, De.85.a.89. % de personas en cada rango de edad. El estudio de estos profiles se puede realizar conjuntamente ya que todos aportan informaciones similares, a medida que aumenta la edad de un municipio aumenta la probabilidad de despoblación, por tanto, para los intervalos de menor edad a medida que aumenta el % disminuye la probabilidad de despoblación y estos gráficos se invierten cuando aumenta el rango de edad.

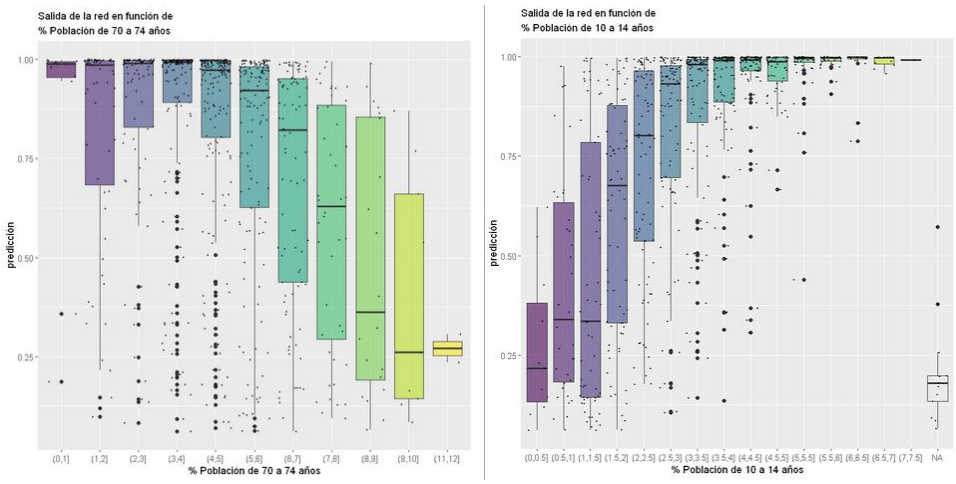


Figura 27. Distribución de la probabilidad de despoblación (siendo 1 no despoblado y 0 despoblado) frente al % de población en un rango de edad de 70 a 74 años, figura de la derecha, y 10 a 14 años, figura de la izquierda.

-Porcentaje Mayor de 65: indica el porcentaje de población mayor de 65 años. La predicción sobre la despoblación tiene una gran dependencia con esta variable disminuyendo la probabilidad de despoblación a medida que disminuye el porcentaje de población mayor a 65 años. Como valor de "turn-on" se puede determinar el 35% de población, a partir de este valor la mayoría de los municipios se predicen como despoblados.

-Porcentaje Menor de 14: indica el porcentaje de población menor de 14 años. La dependencia tiene una forma similar al de la variable de porcentaje mayores de 65 pero de forma inversa. En este caso a medida que la población menor de 14 aumenta disminuye la probabilidad de despoblación. A partir del 8 % la mayoría de los municipios se predicen como poblados.

- Edad media personas: en este caso el nombre de la variable define la medida. La probabilidad de despoblación es aumenta cuanto mayor es la edad media. Como valor de "turn-on" se puede determinar los 50 años de edad, con edades superiores la mayor parte de los municipios se predicen como despoblados.

-**Tamaño medio del hogar**: hace referencia al número medio de personas que viven en un hogar. Gráficamente se puede determinar que en los municipios que se predicen como despoblados hay un menor número de personas residiendo en cada vivienda. Como valor de "turn-on" se puede determinar 2.4 personas por hogar, a partir de este valor los municipios tienen mayor tendencia a ser predichos como poblados.

- **Indicador de emancipación 30 -34:** es el porcentaje de personas entre 30 y 34 años que siguen perteneciendo al núcleo de sus padres respecto al total de personas de ese rango de edad. Para formar parte del núcleo familiar los hijos no deben estar emparejados ni tener hijos. Se observa que a medida que aumenta este indicador aumenta la probabilidad de despoblación, es decir, la emancipación va con retraso en municipios que tienden a la despoblación. A partir del 50% la mayoría de los municipios se predicen como despoblados.

-**Media hijos por núcleo:** es el número de hijos por núcleo familiar entre el número total de núcleos. También tiene una clara dependencia con la salida de la red, a medida que hay mayor número de hijos por núcleo los municipios se predicen como poblados.

- **Nacidos en el municipio:** % de personas respecto del total de población de un municipio que han nacido en él, aunque su lugar de residencia ya no sea dicho municipio. Se observa que a medida que aumenta, aumenta la despoblación.

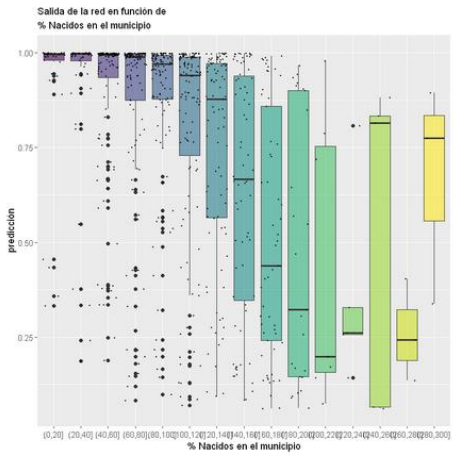


Figura 28. Distribución de la probabilidad de despoblación (siendo 1 no despoblado y 0 despoblado) frente al % de Nacidos en el Municipio.

- **Total viviendas:** número total de viviendas que existe en un municipio en escala logarítmica. El termino vivienda se refiere a cualquier recinto que estaestá concebido para ser habitado por personas. En el profile se observa una tendencia que ha mayor número de viviendas mayor probabilidad de no despoblarse, aunque existen rangos en los que no se sigue la tendencia y para un bajo número de viviendas los municipios se predicen como despoblados.

- **Total locales:** número de locales totales que hay en un municipio. El termino local se refiere a cualquier recinto que no está dedicado a vivienda familiar y en el que se pueden llevar a cabo actividades económicas. Existe una dependencia de la salida de la red con esta variable bastante clara que se puede observar con los locales, donde a medida que aumenta el número aumenta la probabilidad de un municipio de ser predicho como

poblado. Como valor de "turn-on" se puede determinar el 0.5, teniendo en cuenta que para esta variable se ha tomado log(x+1).

- Total edificios: número total de edificios que hay un municipio. Se define edificio como cualquier construcción permanente concebida para ser utilizada como vivienda o para el desarrollo de una actividad. Parece que existe una relación de la salida de la red con esta variable sobre todo para grandes valores ya que estos se predicen siempre como poblados, sin embargo, para valores bajos de esta variable la predicción es más fluctuante.

- Locales.salud: % de locales de salud respecto al total de locales. Se ha representado la variable en escala logarítmica. Es interesante, ya que parece que la probabilidad de despoblación aumenta a medida que aumenta el % de locales de salud.

- Locales.sociales: % de locales sociales respecto al total de locales. La variable se encuentra en escala logarítmica en la representación gráfica. Se observa como, aunque no es muy clara la dependencia, disminuye la probabilidad de despoblación a medida que disminuye el % de locales.

- Viviendas secundarias: vivienda familiar que no es la habitual y se utiliza de forma esporádica. A medida que aumenta el número de viviendas secundarias aumenta la despoblación, aunque se observa una distribución bastante uniforme.

-Antigüedad media edificios: Edad de construcción media del total de edificios Se observa una gran dependencia con la salida de la red, a mayor edad mayor probabilidad de despoblación. Hasta 80 la salida tiene una tendencia a predecir los municipios como poblados y de 80 a 100 como despoblados.

-Indicador de juventud de la población potencialmente activa: proporción de población de 20 a 39 años respecto de la población de 40 a 59 años residente en un municipio.

- Indicador de dependencia juvenil: proporción de población menor de 20 años respecto a la población de 20 a 59 años residente en un municipio. A medida que aumenta disminuye la probabilidad de despoblación, sin embargo, para valores más altos también se observa que predice gran parte de municipios como despoblados.

- Indicador de dependencia senil: porcentaje de población mayor de 59 años respecto a la población entre 20 y 59 años. Como en todas las variables de edad de la población se observa una gran dependencia con la salida de la red. A medida que aumenta este indicador aumenta la posibilidad de que un municipio sea predicho como poblado.

- Antigüedad media en la vivienda: La antigüedad del hogar en la vivienda se obtiene a partir del año de llegada a la misma y, si no llegaron todos los miembros del hogar al mismo tiempo, se refiere al primero que llegó. A medida que aumenta este valor aumenta la probabilidad de despoblación.

- Estudios post-obligatorios: % de personas de 16 a 25 años que están cursando estudios post-obligatorios. En este caso la salida de la red no sigue una tendencia clara en función de la variable, en los primeros rangos disminuye la probabilidad de despoblación con el aumento de la variable, pero para valores altos el comportamiento es inverso.

- Pareja con hijos: % de parejas que residen en un municipio en función del total. A medida que aumenta el número de parejas con hijos disminuye la probabilidad de despoblación, se observa que alrededor del 8 % la mayor parte de los municipios se predicen como poblados, aun así entre 8 y 14 siguen existiendo municipios cuya predicción es la despoblación.

- Dispone de 2º vivienda: porcentaje número de hogares que dispone de una vivienda que se suele usar para vacaciones, fines de semana, etc., entre el número total de hogares. Para valores bajos de esta variable, aunque hay una mayor parte de municipios con una predicción de probabilidad de no estar despoblados alta, también hay un gran número de ellos con predicción de despoblación, lo que se observa más claramente es que a medida que disminuye este valor aumenta la probabilidad de despoblación.

Variables territoriales:

Las variables territoriales ya se detallaron en mayor profundidad en el C**apítulo 2** (histogramas) por tanto se explicarán directamente los profiles obtenidos para las variables seleccionadas.

- Altitud: se observa claramente en el profile la dependencia de la despoblación con la altura. Para el primer rango, de 0 a 200 m, ninguno de los municipios se predice como despoblados y a medida que este valor aumenta la predicción tiene una mayor tendencia a la despoblación.

- Distancia a un núcleo de 10 mil habitantes: inicialmente no se observaba una gran correlación, sin embargo, pasando la variable a escala logarítmica se observa como en los rangos más bajos de esta variable todos los municipios se predicen como poblados y a medida que aumenta la distancia comienza a predecir un mayor número de municipios como despoblados.

- Distancia estación de ferrocarril en metros: aumenta la despoblación a medida que aumenta la distancia. Existe una anomalía para valores muy alto que puede deberse a los municipios de Melilla y las Islas Canarias para los que se alteró este valor.

7. Estudio de la posibilidad de revertir la tendencia poblacional

Existen variables sobre las que no se puede actuar, fundamentalmente estructurales u orográficas, sin embargo, hay otras relacionadas con el trabajo o los estudios sobre las que sabiendo si influyen en la población se pueden intentar realizar cambios y prevenir la despoblación. Se va analizar de forma individualizada cómo el cambio de una variable para cada municipio que se ha predicho como despoblado afecta al resultado de la red, como posible indicación de que actuando sobre esta variable se podría evitar el despoblamiento.

En general los datos obtenidos para todas las redes neuronales estudiadas han sido bastante similares. Para este caso se va a utilizar la red neuronal con la reducción de variables determinada por los profiles, es decir, la red neuronal entrenada con las variables predictoras que se ha determinado que tenían mayor influencia sobre la salida de la red.

Se han seleccionado municipios para los que la red neuronal les proporcionaba la predicción de despoblación y se han alterado las variables con mayor influencia (y que se pueden alterar).

Se ha estudiado de dos formas, de una manera gráfica, representando la probabilidad de despoblación otorgada por la red para una serie de municipios (que inicialmente predecía como despoblados) en función del valor de la variable y de una forma cuantitativa, viendo el porcentaje de municipios que predice como poblados sobre el total de municipios que inicialmente predecía como despoblados. Para la representación gráfica se han cogido 50 municipios que inicialmente se predecían como despoblados ya que al representar todos los municipios se sobrecargaba demasiado el gráfico y era confusa su interpretación.

- **Total 3 (Población residente):** Se han dado valores de 10 a 90 en intervalos de 10 en 10. A medida que el valor de esta variable aumenta disminuye la probabilidad de despoblación. En los profiles se observaba un valor de "turn on" que no se observa ahora gráficamente. Cuantitativamente es a partir de 60 (valor cercano al obtenido en los los profiles), cuando algunos municipios despoblados empiezan a cambiar su predicción.

Valores	% Poblados
0	0
10	0
20	0
30	0
40	0
50	0
60	0.63
70	0.63
80	1.89
90	3.16

100	7.59

Tabla 5. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

- **Solo reside:** Se han dado valores de 10 a 100 en intervalos de 10 en 10. A medida que el valor de esta variable aumenta la probabilidad de despoblación disminuye. En este caso gráficamente se observa una gran correlación y la mayor parte de los municipios seleccionados cambian su predicción significativamente. Cuantitativamente el 6% de los municipios despoblados pasan a poblados cuando el valor es igual a 50 y llegan a pasar el 30 % cuando está en 80.

Valores	% Poblados
0	0
10	0
20	0
30	0.63
40	3.16
50	6.33
60	10.76
70	20.88
80	30.38
90	37.97
100	45.56

Tabla 6. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

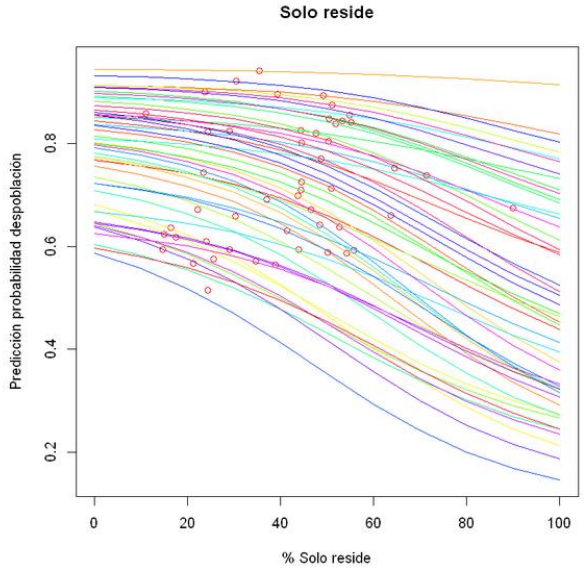


Figura 24. Predicción de despoblación (siendo 1 el máximo valor y 0 el mínimo) frente al % de población que solo reside en un municipio. Las líneas corresponden a los cambios de la predicción en función de la variable y los puntos al valor inicial que tenían estos municipios.

- **Total 8 (% Población vinculada no residente):** en los profiles se observaba que a medida que aumenta este valor aumenta la probabilidad de despoblación. En este caso gráficamente no se observa mucha correlación. Cuantitativamente algunos municipios cambian su predicción, pero para rangos muy bajos de esta variable.

Valores	% Poblados
0	4.04
10	1.26
20	0.63
30	0
40	0
50	0
60	0
70	0
80	0
90	0
100	0

Tabla 7. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

- **Edad media**: aumenta a medida que aumenta la probabilidad de despoblación. Los valores que se han dado han sido dividiendo el rango de edad media de los datos en 9 intervalos iguales. Cuando la edad media tiene un valor de alrededor de 40 años los municipios comienzan a pasar a poblados en la clasificación, corresponde con el valor de turn-on del profile.

Valores	% Poblados
66	0
62	0
59	0
55	0
52	0
48	0
44	0
40	1.89
36	7.59
32	12.02
28	15.18

Tabla 8. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

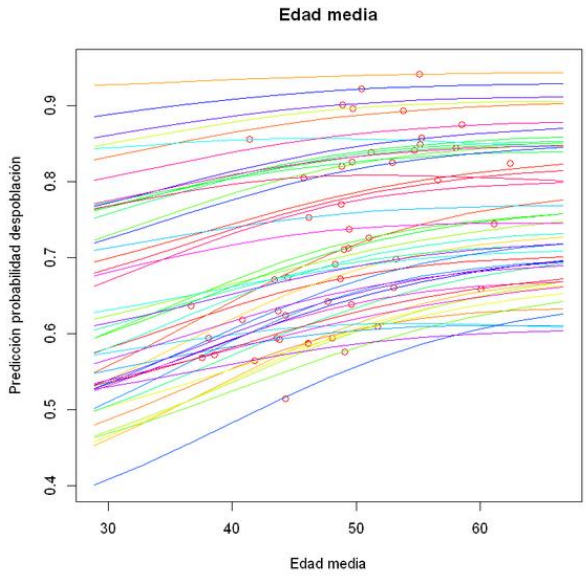


Figura 25. Predicción de despoblación (siendo 1 el máximo valor y 0 el mínimo) frente a la edad media de la población de un municipio. Las líneas corresponden a los cambios de la predicción en función de la variable y los puntos al valor inicial que tenían estos municipios.

- **Distancia ferrocarril:** a medida que disminuye la distancia a una estación de ferrocarril disminuye la probabilidad de despoblación.

Valores	% Poblados
90929	0
81896	0
72864	0
63832	0
54799	0
45767	0
36734	0
27702	0
18670	0.63
9637	1.26
605	5.06

Tabla 9. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

Gráficamente, **Figura 26**, se observa cómo hay municipios a los que el cambio de esta variable les afecta significativamente mientras otros se mantienen estables. Se observa un mayor efecto sobre municipios con probabilidades bajas de despoblación.

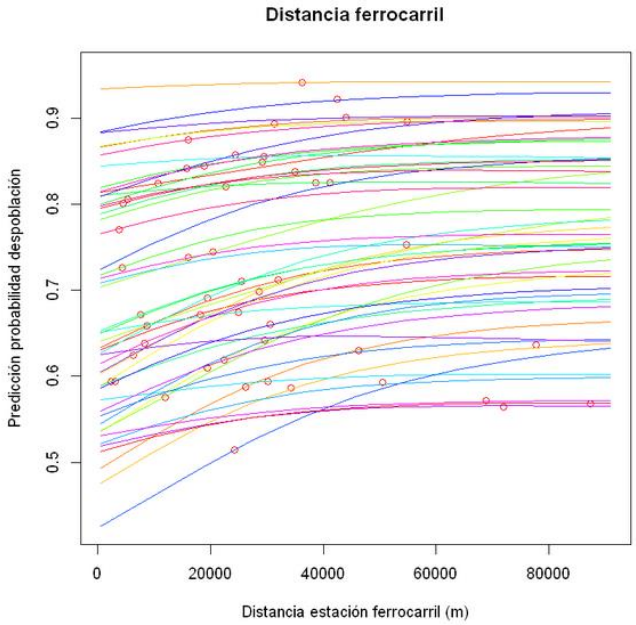


Figura 26. Predicción de despoblación (siendo 1 el máximo valor y 0 el mínimo) frente a la distancia en metros a una estación de ferrocarril. Las líneas corresponden a los cambios de la predicción en función de la variable y los puntos al valor inicial que tenían estos municipios.

- **Estudios post-obligatorios:** a medida que aumenta la gente con estudios post obligatorios disminuye la probabilidad de despoblación. El 12% de los municipios anteriormente predichos como despoblados cambian su predicción cuando el 70% de las personas del municipio tienen estudios post obligatorios.

Valores	% Poblados
0	0
10	0
20	0.63
30	0.63
40	0.63
50	1.89
60	5.06
70	12.02
80	15.82
90	15.82
100	17.72

Tabla 10. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

- **Porque tienen una segunda vivienda**: Se dan valores de 10 a 100 en intervalos de 10 en 10. A medida que va aumentando el % de personas que tienen una segunda vivienda allí aumenta la despoblación. A partir del 50% todos los

municipios se mantienen como despoblados, este valor corresponde con el valor de turn-on que se veía en el profile.

Valores	% Poblados
0	6.96
10	1.26
20	0.63
30	0.63
40	0.63
50	0
60	0
70	0
80	0
90	0
100	0

Tabla 11. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

- **Porque trabaja allí**: Se dan valores de 10 a 100 en intervalos de 10 en 10. A medida que aumenta la gente que reside en un municipio porque trabaja allí disminuye la probabilidad de despoblación, a partir del 50% un 15% de los despoblados han cambiado su predicción.

Valores	% Poblados
0	0
10	0.63
20	1.89
30	9.49
40	12.65
50	15.18
60	15.82
70	16.45
80	17.08
90	17.72
100	20.25

Tabla 12. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

Gráficamente se observa la diferencia en el cambio de predicción de unos municipios a otros, aunque la predicción disminuye o se mantiene estable para todos los casos para algunos municipios este descenso está mucho más marcado. Es interesante observar también como todos los municipios tenían inicialmente valores muy bajos, cercanos a 0, de porcentaje de población vinculada porque que trabaja en ese municipio.

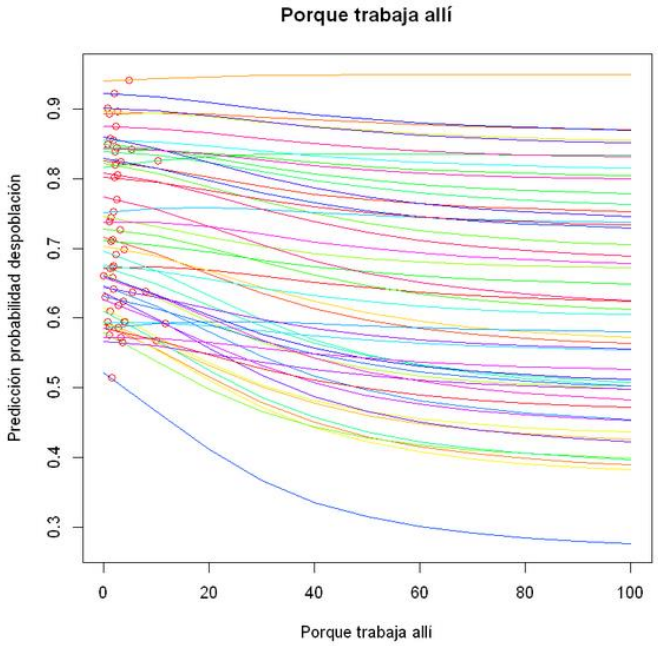


Figura 27. Predicción de despoblación (siendo 1 el máximo valor y 0 el mínimo) frente al % de población que está vinculada al municipio porque trabaja allí. Las líneas corresponden a los cambios de la predicción en función de la variable y los puntos al valor inicial que tenían estos municipios.

- **Índice de dependencia juvenil**: a medida que aumenta disminuye la probabilidad de despoblación, es a partir de 0.4 cuando municipios despoblados cambian su predicción, en cambio en el profile el valor de turn-on era inferior.

Valores	% Poblados
0	0
0.1	0
0.2	0
0.3	0
0.4	1.89
0.5	3.79
0.6	8.23
0.7	10.76
0.8	12.66
0.9	14.56
1.0	16.45

Tabla 13. En esta tabla se encuentran los valores, en la primera columna, de la variable y el correspondiente tanto por ciento de municipios despoblados que cambiaban su predicción a poblados, segunda columna.

Es interesante la comparación entre los profiles, gráficos que mostraban la dependencia de la predicción y las variables, y el cambio de predicción al alterar las variables. Aunque en algunos la variación de la predicción de los municipios sigue la misma tendencia observada en los profiles en otros no, como por ejemplo en Total8 (% Población vinculada no residente), donde, aun siendo una de las variables seleccionadas por tener mayor dependencia no se observaba casi ningún cambio de predicción.

Esto refleja que la red neuronal realiza las predicciones a partir de la combinación de las variables y puede que sea una relación junto con otra variable la que afecte a la salida. Esto se observa también con las variables Total 3 (% Población residente), Emancipación y Locales de salud.

De las variables estudiadas con este método se puede determinar que hay una clara dependencia con la edad en la predicción de despoblación de un municipio. Esta variable puede ser una consecuencia de la despoblación más que un motivo de ella y resulta difícil actuar directamente sobre ella. Sin embargo, otras variables como el estudio y el trabajo también son influyentes a la hora de determinar la despoblación y en estos campos sí que se pueden aplicar medidas que puedan favorecer a la creación de empleo. De hecho, el empleo puede ser consecuencia de los estudios. Fomentar la educación ayudará en un futuro a crear empleo. Empleo y estudio ayuda también a tener una población más joven lo que se ve que afecta mucho a la despoblación.

8. Conclusiones

Con la realización de este trabajo se pretendía abordar un problema del ámbito demográfico a partir de técnicas de Machine Learning, aportando un enfoque novedoso a este campo que pueda proporcionar interesantes resultados.

Los datos a los que se tenía acceso y que se han utilizado para entrenar el algoritmo han sido los censos de los años 2001 y 2011 desarrollados por el Instituto Nacional de Estadística y las variables territoriales, obtenidas por análisis SIG.

Tras probar distintos modelos de redes neuronales se ha determinado que el que proporciona mejores resultados en cuanto a la predicción de despoblación es aquel que es entrenado con un dataset constituido por 46 variables predictoras, 1 variable objetico y con municipios de poblaciones inferiores a 50.000 habitantes. La red estaba formada por 2 capas ocultas de 20 neuronas la primera y 2 la de salida. Este modelo ha obtenido un valor de función de coste para la muestra de validación **fcoste** = 0.0813 (**Tabla 4**).

Hay que tener en cuenta que en este modelo se han eliminado los municipios con grandes poblaciones que constituían el 2.5 % del total que son municipios que la red predeciría correctamente al ser casos con muy poca probabilidad de despoblación. Se ha obtenido una capacidad de clasificar como verdaderos positivos a más del 90% de los municipios, considerando casos positivos los municipios poblados y negativos los despoblados, y como verdaderos negativos alrededor del 70%.

A partir del análisis de las variables y la distribución de los municipios se ha determinado que los municipios despoblados se concentran en zonas de interior y de montaña a lo largo del territorio nacional, que la despoblación en 2011 surge en zonas colindantes con los municipios despoblados en el periodo anterior y que los indicadores de edad son aquellos en los que más se observa la diferencia de distribución en función de la ocurrencia de despoblación.

Con el estudio de la posibilidad de revertir la tendencia poblacional se ha confirmado que las variables de edad son unas de las más influyentes sobre el fenómeno de la despoblación, pero además se ha observado gran cambio al actuar sobre indicadores del trabajo y estudios post-obligatorios.

Este trabajo ha obtenido resultados satisfactorios en cuanto a la predicción de un fenómeno tan complejo como la despoblación, que depende de numerosas variables de distintos ámbitos. Así como en el estudio de la dependencia de las variables donde, se ha confirmado que el envejecimiento de la población es uno de los factores que más afecta a la despoblación, y que variables como la existencia de segundas viviendas o la vinculación por motivos de trabajo también son indicadores de este fenómeno.

La aplicación de técnicas de aprendizaje automático es de gran interés en este campo, ya que se ha determinado que obtienen resultados prometedores y pueden ser una herramienta muy útil para futuros trabajos.

Bibliografía

[1] Aston Zhang, Zachary C. Lipton, Mu Li and Alexander J.Smola. (2020) Dive into Deep Learning.

[2] E. Castillo, A. Cobo, J.M. Gutiérrez, and E. Pruneda (1999) Introduction to Functional Networks with Applications. A Neural Based Paradigm. Springer International Series in Engineering and Computer Science. Chapters 1 and 2.

[3] Olga de Cos Guerra, Pedro Reques Velasco. Vulnerabilidad territorial y demográfica en España. Posibilidades del análisis multicriterio y la lógica difusa para la definición de patrones espaciales. Investigaciones Regionales – Journal of Regional Research, 45 (2019/3). ISSN: 1695-7253.

[4] Shai Shalev-Schwartz, Shai Ben-David. Undestanding Machine Learning. From theroy to algorithms. Cambridge University Press. ISBN 978-1-107-05713-5

[5]http://www.femp.es/sites/default/files/multimedia/documento_de_accion_comision_de_despoblacion_9-05-17.pdf

[6] Real Decreto 40/2017, de 27 de enero, por el que se crea el Comisionado del Gobierno frente al Reto Demográfico y se regula su régimen de funcionamiento. «BOE» núm. 24, de 28 de enero de 2017. Referencia: BOE-A-2017-915

[7] Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

[8] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[9] J Allaire and François Chollet (2020). keras: R Interface to 'Keras'. R package version 2.3.0.0. https://CRAN.R-project.org/package=keras

[10] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[11] https://www.ine.es/

[12] Ministerio de Fomento; Dg de Arquitectura, Vivienda y Suelo; Información y Evaluación. Áreas urbanas en España 2018. Constitución, Cuarenta años de las ciudades españolas. Pg 20

[13] Bosque Sedra, Joaquin. Espacio geográfico y ciencias sociales. Nuevas propouestas para el estudio del territorio. Investigaciones Regionales, núm. 6, 2005, pp.103-11º. Asociación Española de Ciencia Regional.

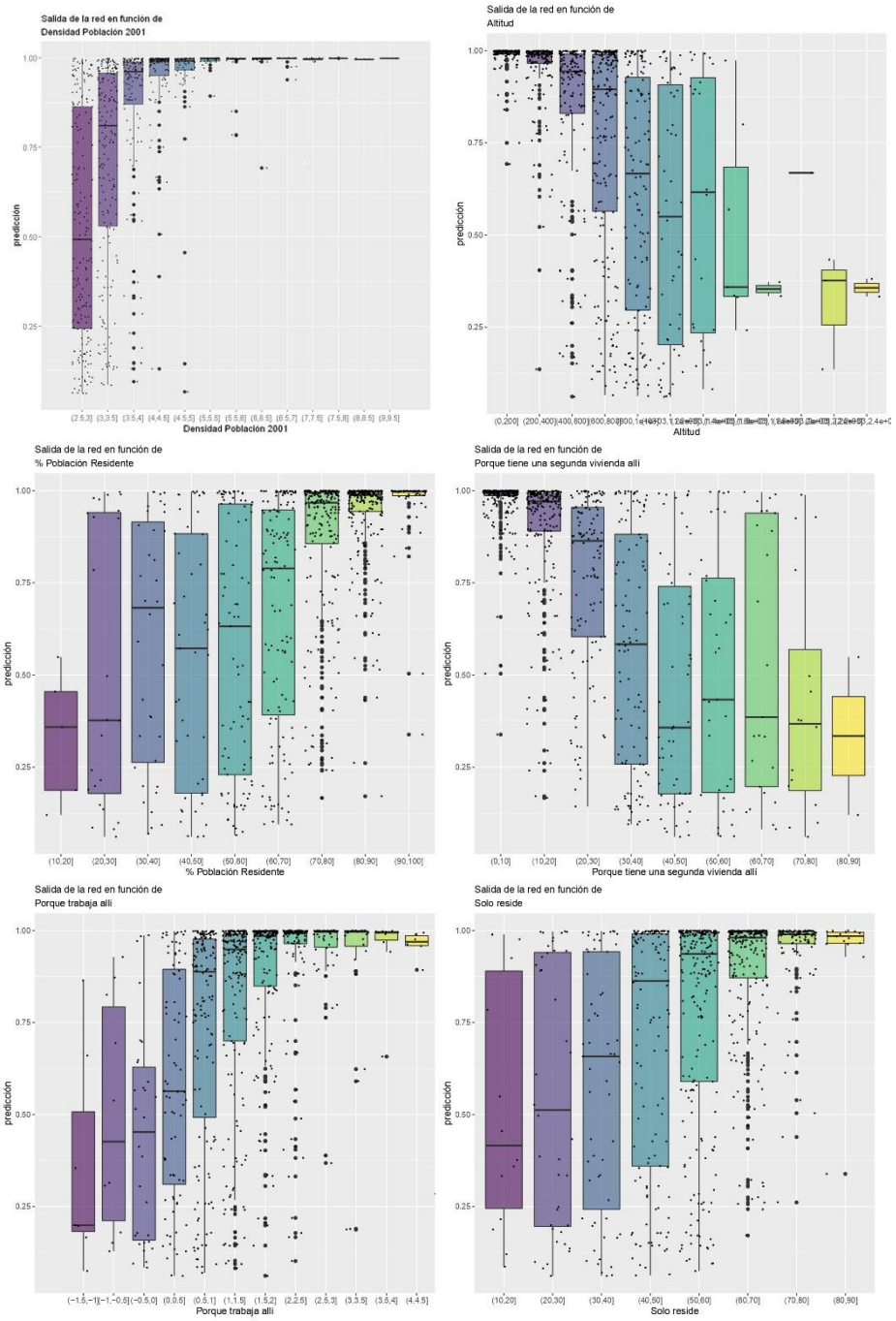
[14] Haider Khalaf Jabbar, Dr. Rafiqul Zaman Khan. Methods to avoid over-fitting and under-fitting in supervised machine learning. ISBN: 978-981-09-5246-1. Pg (164-165).

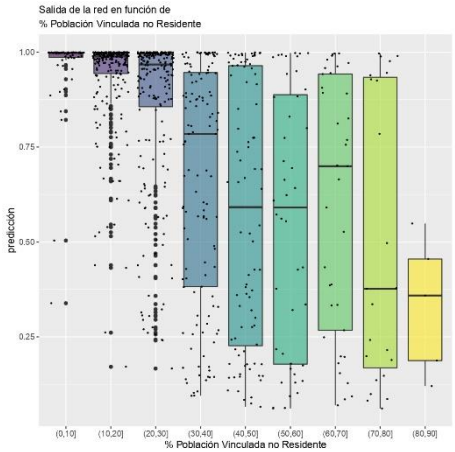
[15] https://towardsdatascience.com/what-are-overfitting-and-underfitting-in-machine-learning-a96b30864690

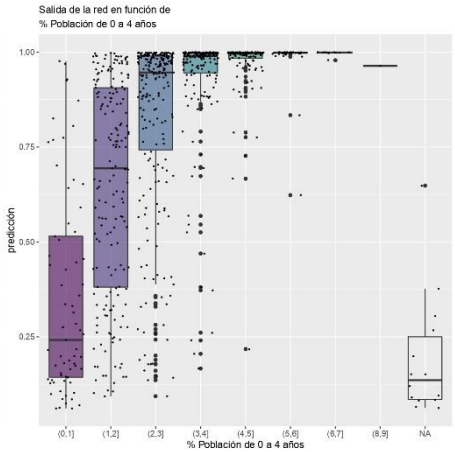
[16] IGN, Base cartográfica de límites administrativos municipales e INE, Censos de población 2001 y 2011.

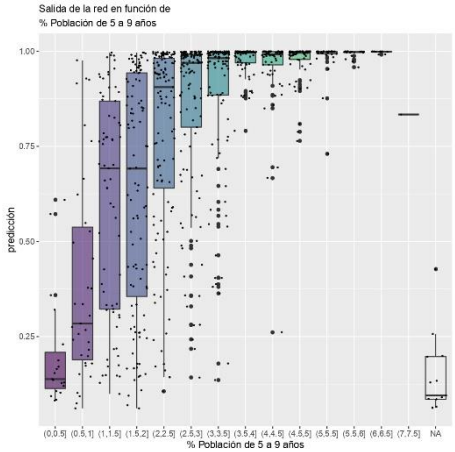
Anexo

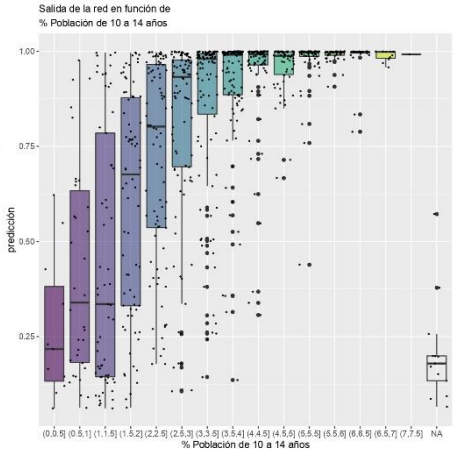
- Profiles para el modelo seleccionado

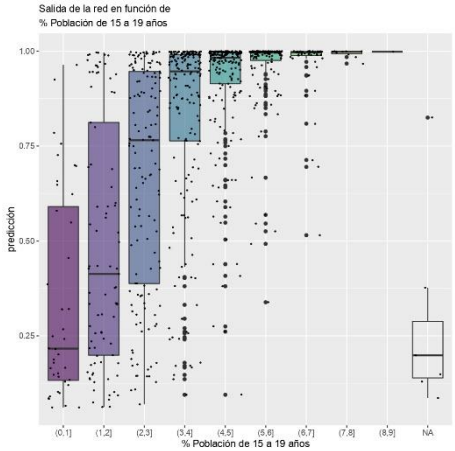


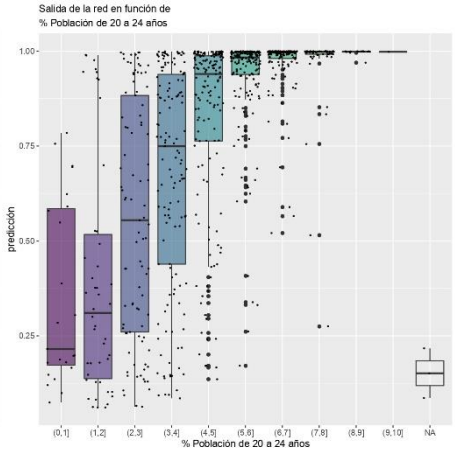

Salida de la red en función de
% Población Vinculada no Residente
predicción
1.00
0.75
0.50
0.25
(0,10]
(10,20]
(20,30]
(30,40]
(40,50]
(50,60]
(60,70]
(70,80]
(80,90]
% Población Vinculada no Residente

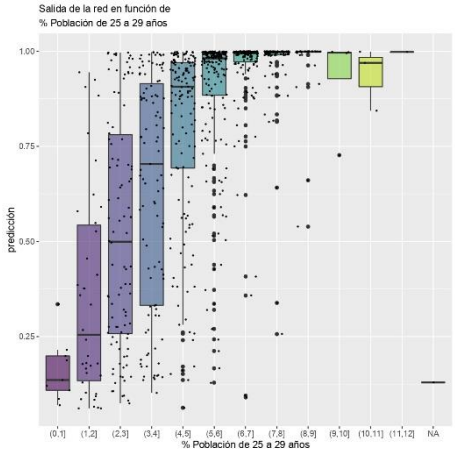

Salida de la red en función de
% Población de 0 a 4 años
predicción
1.00
0.75
0.50
0.25
(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(8,9]
NA
% Población de 0 a 4 años

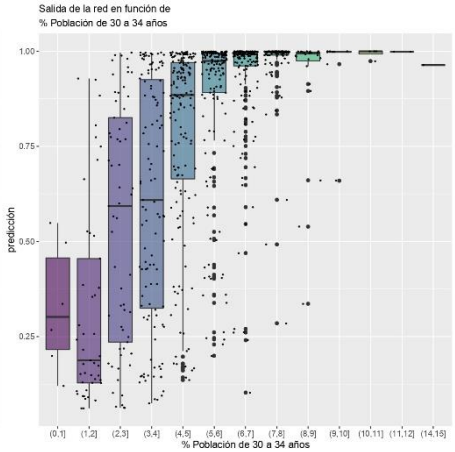

Salida de la red en función de
% Población de 5 a 9 años
predicción
1.00
0.75
0.50
0.25
(0,0.5]
(0.5,1]
(1,1.5]
(1.5,2]
(2,2.5]
(2.5,3]
(3,3.5]
(3.5,4]
(4,4.5]
(4.5,5]
(5,5.5]
(5.5,6]
(6,6.5]
(7,7.5]
NA
% Población de 5 a 9 años

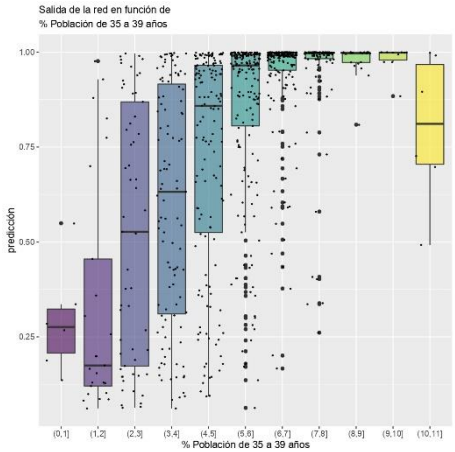

Salida de la red en función de
% Población de 10 a 14 años
predicción
1.00
0.75
0.50
0.25
(0,0.5]
(0.5,1]
(1,1.5]
(1.5,2]
(2,2.5]
(2.5,3]
(3,3.5]
(3.5,4]
(4,4.5]
(4.5,5]
(5,5.5]
(5.5,6]
(6,6.5]
(7,7.5]
NA
% Población de 10 a 14 años

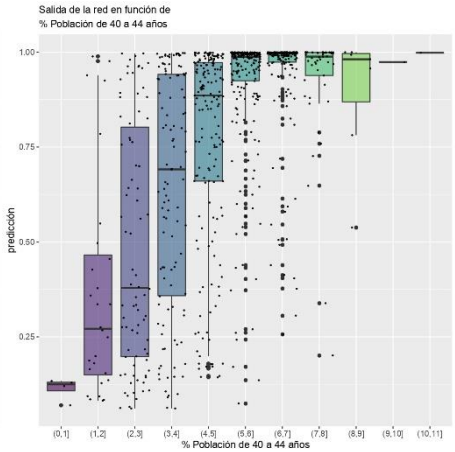

Salida de la red en función de
% Población de 15 a 19 años
predicción
1.00
0.75
0.50
0.25
(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]
(8,9]
NA
% Población de 15 a 19 años

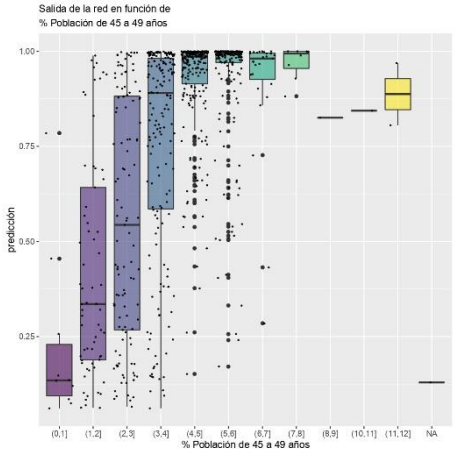

Salida de la red en función de
% Población de 20 a 24 años
predicción
1.00
0.75
0.50
0.25
(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]
(8,9]
(9,10]
NA
% Población de 20 a 24 años

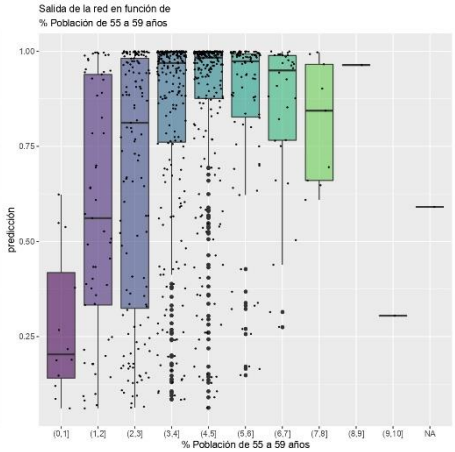

Salida de la red en función de
% Población de 25 a 29 años
predicción
1.00
0.75
0.50
0.25
(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]
(8,9]
(9,10]
(10,11]
(11,12]
NA
% Población de 25 a 29 años

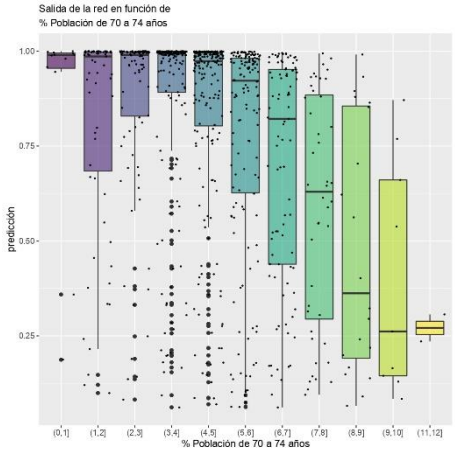

Salida de la red en función de
% Población de 30 a 34 años
predicción
1.00
0.75
0.50
0.25
(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]
(8,9]
(9,10]
(10,11]
(11,12]
(14,15]
% Población de 30 a 34 años

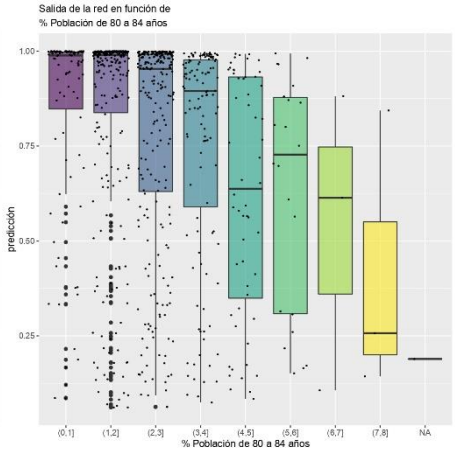

Salida de la red en función de
% Población de 35 a 39 años
predicción
1.00
0.75
0.50
0.25
(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] (8,9] (9,10] (10,11]
% Población de 35 a 39 años

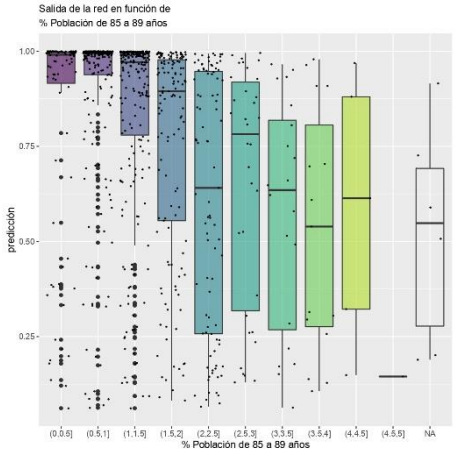

Salida de la red en función de
% Población de 40 a 44 años
predicción
1.00
0.75
0.50
0.25
(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] (8,9] (9,10] (10,11]
% Población de 40 a 44 años

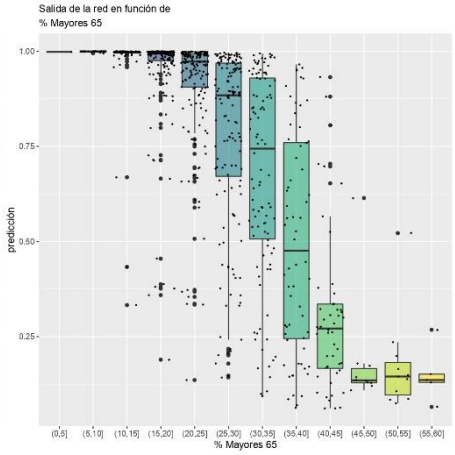

Salida de la red en función de
% Población de 45 a 49 años
predicción
1.00
0.75
0.50
0.25
(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] (8,9] (10,11] (11,12] NA
% Población de 45 a 49 años

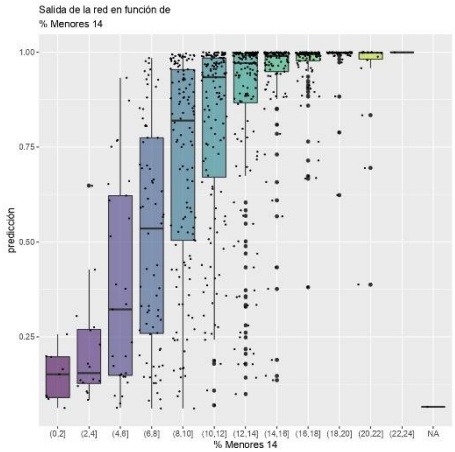

Salida de la red en función de
% Población de 55 a 59 años
predicción
1.00
0.75
0.50
0.25
(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] (8,9] (9,10] NA
% Población de 55 a 59 años

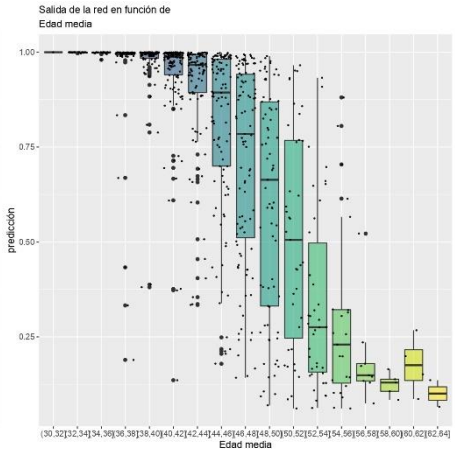

Salida de la red en función de
% Población de 70 a 74 años
predicción
1.00
0.75
0.50
0.25
(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] (8,9] (9,10] (11,12]
% Población de 70 a 74 años

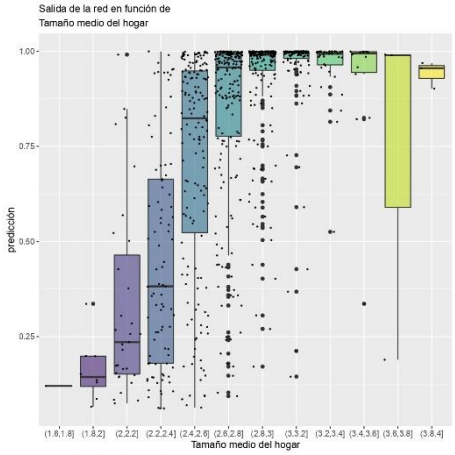

Salida de la red en función de
% Población de 80 a 84 años
predicción
1.00
0.75
0.50
0.25
(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] NA
% Población de 80 a 84 años

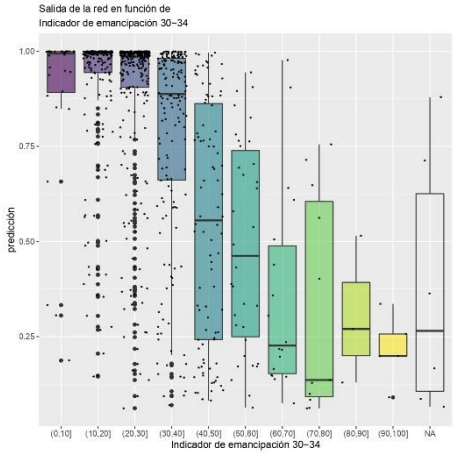

Salida de la red en función de
% Población de 85 a 89 años
predicción
1.00
0.75
0.50
0.25
(0,0.5] (0.5,1] (1,1.5] (1.5,2] (2,2.5] (2.5,3] (3,3.5] (3.5,4] (4,4.5] (4.5,5] NA
% Población de 85 a 89 años

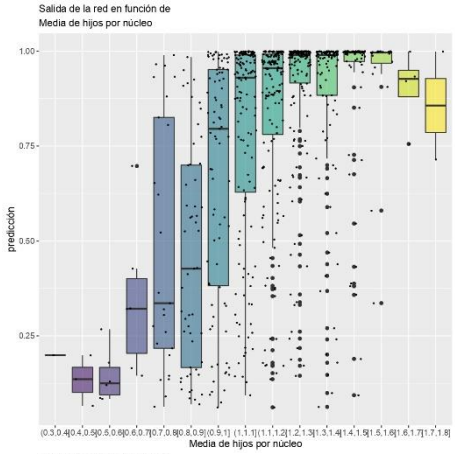

Salida de la red en función de
% Mayores 65
predicción
1.00
0.75
0.50
0.25
(0,5] (5,10] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60]
% Mayores 65

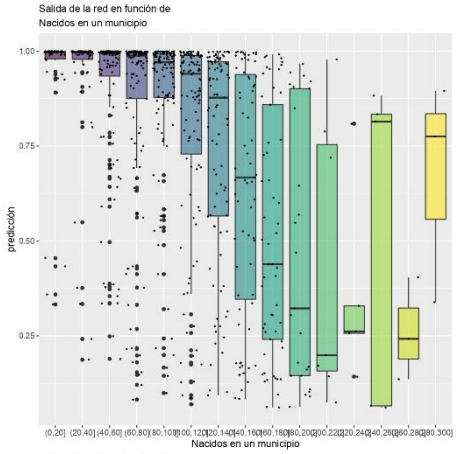

Salida de la red en función de
% Menores 14
predicción
1.00
0.75
0.50
0.25
(0,2] (2,4] (4,6] (6,8] (8,10] (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] NA
% Menores 14

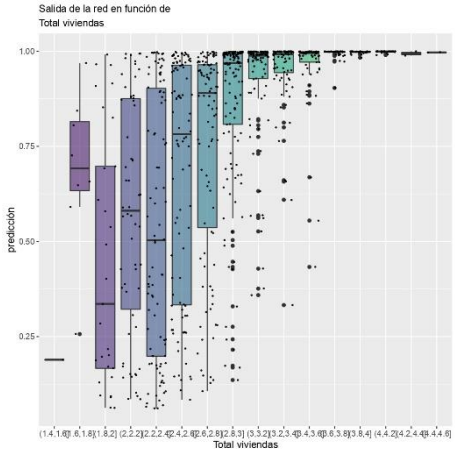

Salida de la red en función de
Edad media
predicción
1.00
0.75
0.50
0.25
(30,32] (32,34] (34,36] (36,38] (38,40] (40,42] (42,44] (44,46] (46,48] (48,50] (50,52] (52,54] (54,56] (56,58] (58,60] (60,62] (62,64]
Edad media

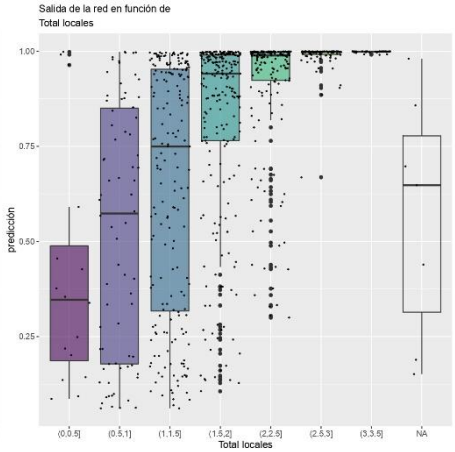

Salida de la red en función de
Tamaño medio del hogar
predicción
1.00
0.75
0.50
0.25
(1.8,1.8] (1.8,2] (2,2.2] (2.2,2.4] (2.4,2.6] (2.6,2.8] (2.8,3] (3,3.2] (3.2,3.4] (3.4,3.6] (3.6,3.8] (3.8,4]
Tamaño medio del hogar

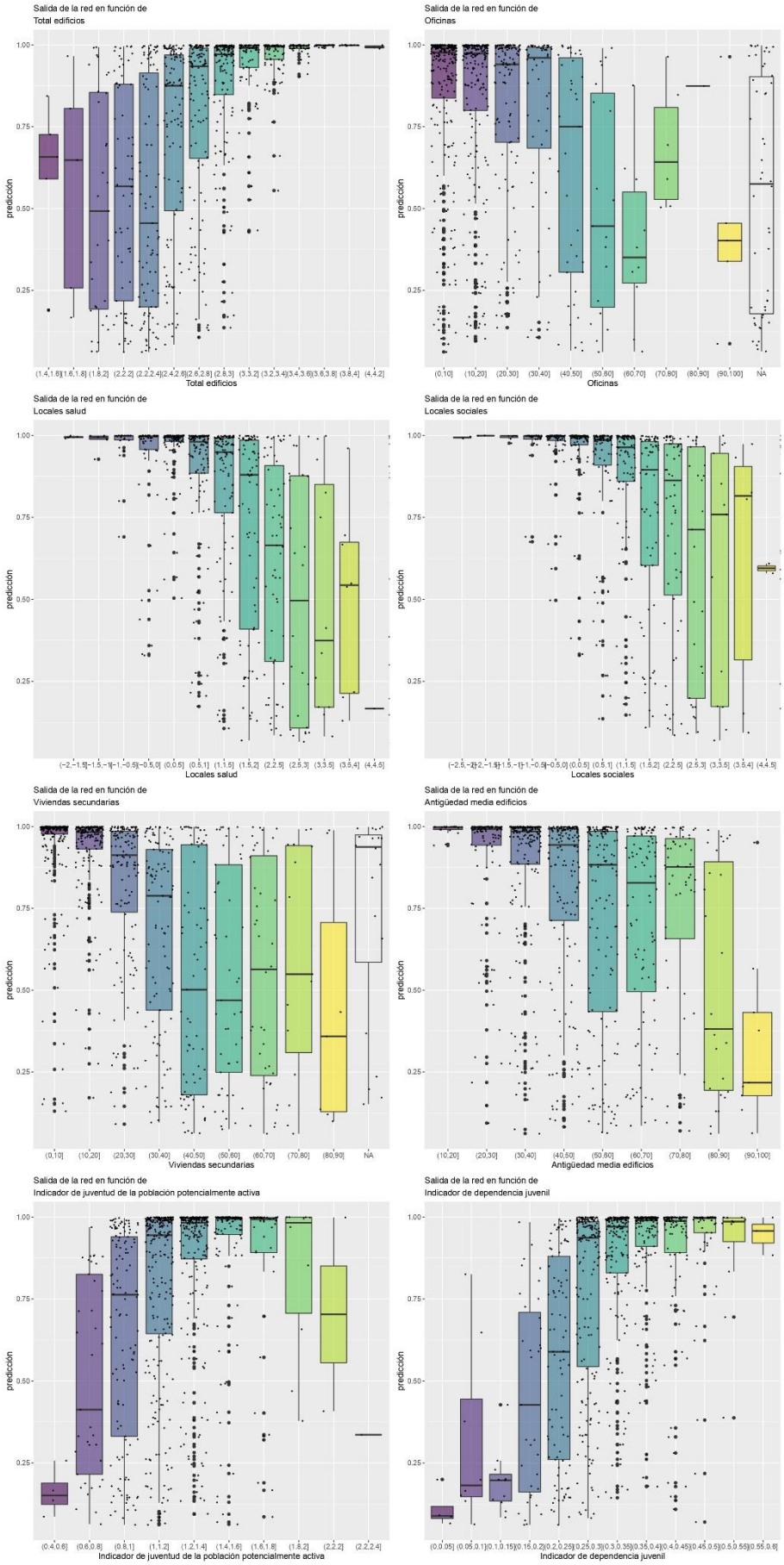

Salida de la red en función de
Indicador de emancipación 30−34
predicción
1.00
0.75
0.50
0.25
(0,10] (10,20] (20,30] (30,40] (40,50] (50,60] (60,70] (70,80] (80,90] (90,100] NA
Indicador de emancipación 30−34

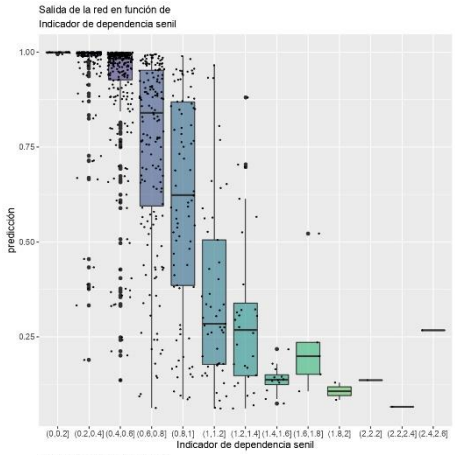

Salida de la red en función de
Media de hijos por núcleo
predicción
1.00
0.75
0.50
0.25
(0.3,0.4] (0.4,0.5] (0.5,0.6] (0.6,0.7] (0.7,0.8] (0.8,0.9] (0.9,1] (1,1.1] (1.1,1.2] (1.2,1.3] (1.3,1.4] (1.4,1.5] (1.5,1.6] (1.6,1.7] (1.7,1.8]
Media de hijos por núcleo

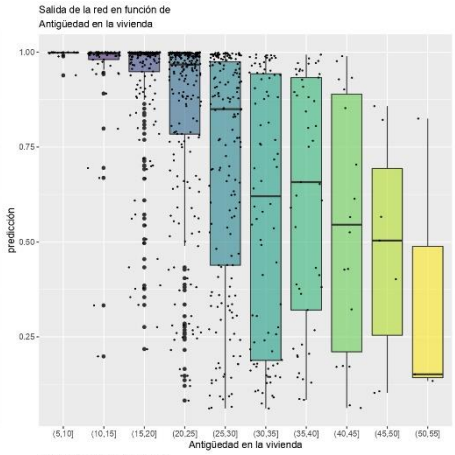

Salida de la red en función de
Nacidos en un municipio
predicción
1.00
0.75
0.50
0.25
(0,20] (20,40] (40,60] (60,80] (80,100] (100,120] (120,140] (140,160] (160,180] (180,200] (200,220] (220,240] (240,260] (260,280] (280,300]
Nacidos en un municipio

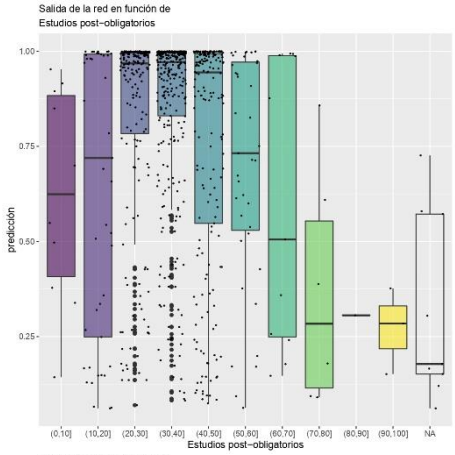

Salida de la red en función de
Total viviendas
predicción
1.00
0.75
0.50
0.25
(1.4,1.6] (1.6,1.8] (1.8,2] (2,2.2] (2.2,2.4] (2.4,2.6] (2.6,2.8] (2.8,3] (3,3.2] (3.2,3.4] (3.4,3.6] (3.6,3.8] (3.8,4] (4,4.2] (4.2,4.4] (4.4,4.6]
Total viviendas

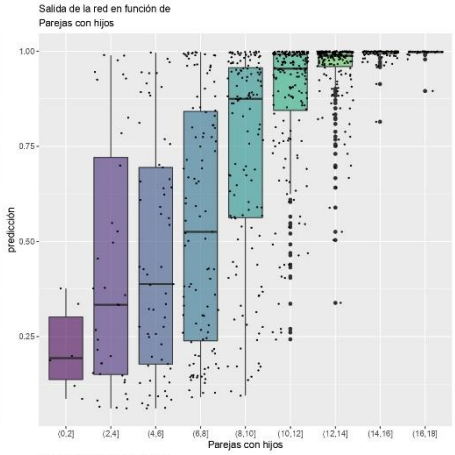

Salida de la red en función de
Total locales
predicción
1.00
0.75
0.50
0.25
(0,0.5] (0.5,1] (1,1.5] (1.5,2] (2,2.5] (2.5,3] (3,3.5] NA
Total locales

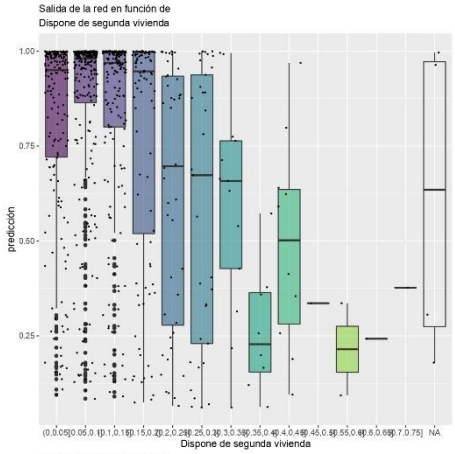

Salida de la red en función de
Total edificios
predicción
Total edificios
Salida de la red en función de
Oficinas
predicción
Oficinas
Salida de la red en función de
Locales salud
predicción
Locales salud
Salida de la red en función de
Locales sociales
predicción
Locales sociales
Salida de la red en función de
Viviendas secundarias
predicción
Viviendas secundarias
Salida de la red en función de
Antigüedad media edificios
predicción
Antigüedad media edificios
Salida de la red en función de
Indicador de juventud de la población potencialmente activa
predicción
Indicador de juventud de la población potencialmente activa
Salida de la red en función de
Indicador de dependencia juvenil
predicción
Indicador de dependencia juvenil

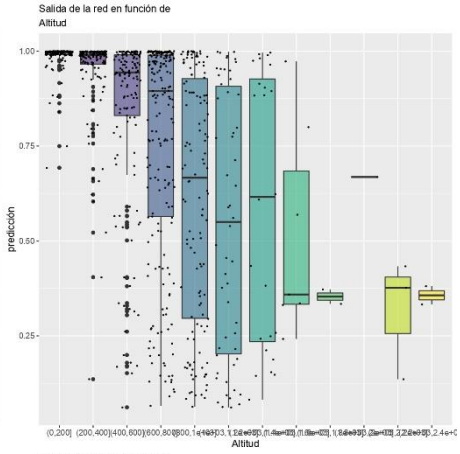

Salida de la red en función de
Indicador de dependencia senil
predicción
1.00
0.75
0.50
0.25
(0,0.2] (0.2,0.4] (0.4,0.6] (0.6,0.8] (0.8,1] (1,1.2] (1.2,1.4] (1.4,1.6] (1.6,1.8] (1.8,2] (2,2.2] (2.2,2.4] (2.4,2.6]
Indicador de dependencia senil

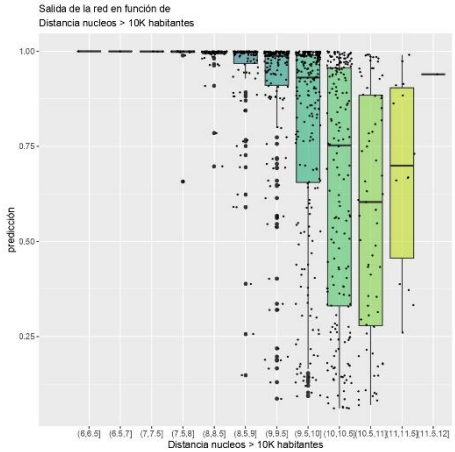

Salida de la red en función de
Antigüedad en la vivienda
predicción
1.00
0.75
0.50
0.25
(5,10] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55]
Antigüedad en la vivienda

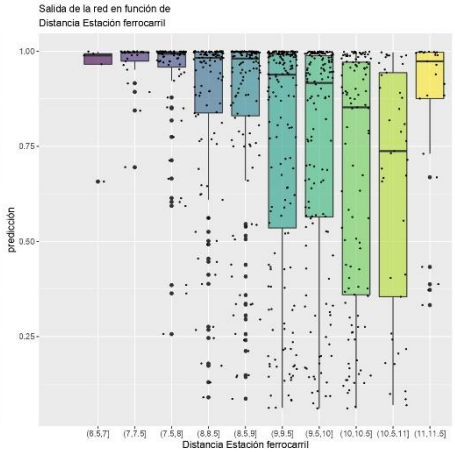

Salida de la red en función de
Estudios post−obligatorios
predicción
1.00
0.75
0.50
0.25
(0,10] (10,20] (20,30] (30,40] (40,50] (50,60] (60,70] (70,80] (80,90] (90,100] NA
Estudios post−obligatorios


Salida de la red en función de
Parejas con hijos
predicción
1.00
0.75
0.50
0.25
(0,2] (2,4] (4,6] (6,8] (8,10] (10,12] (12,14] (14,16] (16,18]
Parejas con hijos


Salida de la red en función de
Dispone de segunda vivienda
predicción
1.00
0.75
0.50
0.25
Dispone de segunda vivienda


Salida de la red en función de
Altitud
predicción
1.00
0.75
0.50
0.25
(0,200] (200,400] (400,600]
Altitud


Salida de la red en función de
Distancia nucleos > 10K habitantes
predicción
1.00
0.75
0.50
0.25
(6,6.5] (6.5,7] (7,7.5] (7.5,8] (8,8.5] (8.5,9] (9,9.5] (9.5,10] (10,10.5] (10.5,11] (11,11.5] (11.5,12]
Distancia nucleos > 10K habitantes


Salida de la red en función de
Distancia Estación ferrocarril
predicción
1.00
0.75
0.50
0.25
(6,6.5] (6.5,7] (7,7.5] (7.5,8] (8,8.5] (8.5,9] (9,9.5] (9.5,10] (10,10.5] (10.5,11] (11,11.5]
Distancia Estación ferrocarril