**CUADERNOS DE TRABAJO**

**FACULTAD DE ESTUDIOS ESTADÍSTICOS**

# Estimación de la tasa de retorno de la carta del censo de los Estados Unidos a través del modelo de regresión lineal y técnicas de predicción inteligentes.

José Luis Jiménez-Moro
Javier Portela García Miguel

***Cuaderno de Trabajo número 01/2014***


UCM

UNIVERSIDAD
COMPLUTENSE
MADRID

Los Cuadernos de Trabajo de la Facultad de Estudios Estadísticos constituyen una apuesta por la publicación de los trabajos en curso y de los informes técnicos desarrollados desde la Facultad para servir de apoyo tanto a la docencia como a la investigación.

Los Cuadernos de Trabajo se pueden descargar de la página de la Biblioteca de la Facultad www.ucm.es/BUCM/est/ y en la sección de investigación de la página del centro www.ucm.es/centros/webs/eest/

CONTACTO:

Biblioteca de la Facultad de Estudios Estadísticos

Universidad Complutense de Madrid

Av. Puerta de Hierro, S/N

28040 Madrid

Tlf. 913944035

buc_est@buc.ucm.es

# Estimación de la tasa de retorno de la carta del censo de los Estados Unidos a través del modelo de regresión lineal y técnicas de predicción inteligentes.

José Luis Jiménez-Moro, Javier Portela García-Miguel
Universidad Complutense de Madrid

**Resumen**

En este proyecto se ha realizado un estudio de la tasa de respuesta que obtiene el censo de los Estados Unidos con la carta que envía a sus ciudadanos. En el estudio se han utilizado 3 técnicas de predicción distintas valorando el error cuadrático medio en cada una de ellas. Los resultados más precisos se han obtenido con una *Red Neuronal*, seguido del algoritmo *Gradient Boosting* y el modelo de regresión lineal. La forma de valorar la precisión de las técnicas predictivas se ha basado no solo en el cálculo del error cuadrático medio sino que además, éste ha sido representado sobre un mapa del país permitiendo ver la mejora en precisión en función de la técnica utilizada.

# 1. Introducción

## 1.1. Motivación

Cada vez vivimos en un mundo más interconectado. Por la mañana, Google sabe donde estoy, Facebook me pide que actualize mi estado y Twitter registra que voy a llegar tarde a clase porque los autobuses están #ApoyandoLaHuelga. Además, al encender el ordenador me aparecen ofertas relacionadas con mi última escapada. ¿Cómo es esto posible? Todos estos datos que se generan día a día son procesados para buscar patrones o comportamientos que sirven no sólo para recomendar un libro que leer o un sitio al que ir sino para realizar predicciones financieras, hacer una estimación de las ventas de una empresa o saber si un e-mail que nos llega es o no spam.

Vivimos en el mundo del *Big Data*, y en este trabajo, que supone la finalización del Grado en Estadística Aplicada, se presenta el funcionamiento de algunas técnicas de predicción apropiadas para procesar grandes volúmenes de información vistas en general a lo largo del Grado y, en particular, en la asignatura "Técnicas Avanzadas de Predicción", impartida por el profesor y también tutor de este trabajo, Javier Portela.

## 1.2. Descripción del trabajo

Los datos de este proyecto provienen de la oficina del censo de los Estados Unidos que, a través de Kaggle, propuso una competición en la que se premiaba el mejor modelo de predicción de la tasa de retorno de las cartas que manda la oficina del censo a los ciudadanos norteamericanos. Se dispone de 71 variables y de 129.605 observaciones.

Kaggle es una página web donde cualquier empresa u organización puede proponer un concurso para obtener el mejor modelo de predicción.

Este trabajo consta de varias partes. En la sección 1, se introduce la motivación con la que se ha llevado a cabo este proyecto. En la sección 2, se detallan los objetivos que se pretenden alcanzar, así como una breve explicación de la metodología empleada. En la sección 3, se explican los fundamentos teóricos de las técnicas de predicción utilizadas. En la sección 4, se muestran y comentan los resultados y finalmente, en la sección 5, se exponen las conclusiones obtenidas.

# 2. Objetivos y metodología

El objetivo principal de este trabajo es obtener un modelo de regresión lineal que cometa el menor error posible a la hora de llevar a cabo la estimación de la tasa de retorno de la carta que la oficina del Censo de los Estados Unidos envía a cada uno de sus ciudadados, y que debe ser cumplimentada y devuelta.

El objetivo secundario es ver si, mediante técnicas de aprendizaje automático o inteligentes, se puede disminuir el error que comete el modelo de regresión lineal.

La metodología empleada en este trabajo se puede dividir en cuatro fases principales:

- **Depuración de datos:** Se ha realizado la depuración de la base de datos donde la principal tarea era imputar los valores perdidos o *missing* en las variables que así lo han requerido. Para la imputación se han utilizando estimaciones censales provistas por la oficina del censo de los Estados Unidos.
- **Análisis descriptivo y transformación de variables:** Se ha efectuado un análisis descriptivo de todas las variables que componen la muestra. Este análisis ha tenido por objeto el estudio de la distribución de cada variable y ver qué transformación era la más adecuada para normalizar su distribución, en el caso de que fuese necesario.
- **Selección de variables:** Una vez obtenidas las variables depuradas y transformadas, se ha buscado un modelo de regresión lineal que ha servido para tener una primera estimación de la tasa de retorno de las cartas del Censo de Estados Unidos.
- **Utilización de técnicas de aprendizade automático o inteligentes para la mejora de la predicción:** Tras tener una primera estimación de la tasa de retorno, se ha tratado de ver si el uso de técnicas de aprendizaje automático o inteligentes mejoraba el resultado que se obtiene mediante un modelo de regresión lineal.

# 3. Fundamentos teóricos

Según la metodología explicada en la sección 2, ha sido necesario realizar una depuración de los datos y transformar las variables que así lo han requerido antes de construir un modelo de predicción. Para implementar la depuración de los datos se han utilizado estimaciones censales provistas por la oficina del Censo de los Estados Unidos, donde ha habido que sustituir cada valor *missing* por su estimación censal correspondiente.

Una vez terminado el proceso de imputación, se han realizado las transformaciones de las variables, tanto de la variable dependiente como de las variables independientes. Para ello, se han elaborarado distintos histogramas que han permitido ver la distribución de cada una de las variables. Con esta técnica se han tratado de identificar las transformaciones las más adecuadas para normalizar sus distribuciones, y conseguir la relación más lineal posible entre la variable dependiente y las variables independientes. Además, se han realizado 2 diagramas de dispersión sobre 4 variables con los que se ve la relación entre la varible dependiente y la variable independiente en su estado original, y la relación entre la varible dependiente y la variable independiente tras haber realizado la transformación de sus observaciones.

A la hora de valorar la distribución de una variable, según se muestra en la Fig.1, existen varias transformaciones posibles en función del tipo de asimetría que se posea. En casos de asimetría positiva, donde la distribución presenta una cola derecha más larga que la izquierda, se ha optado por tomar logaritmos neperianos o la raíz cuadrada de las observaciones de la variable. Por el contrario, ante casos de asimetría negativa, donde la distribución presenta una cola izquierda más larga que la derecha, se han tomado también logaritmos neperianos o la raíz cuadrada de las observaciones de la variables, sólo que reflejando el valor antes de realizar la transformación. Señalamos que para reflejar el valor, por ejemplo, de la variable $y$, hay que buscar su máximo valor $(K)$ y operar con $K - y$, para posteriormente aplicar la transformación.

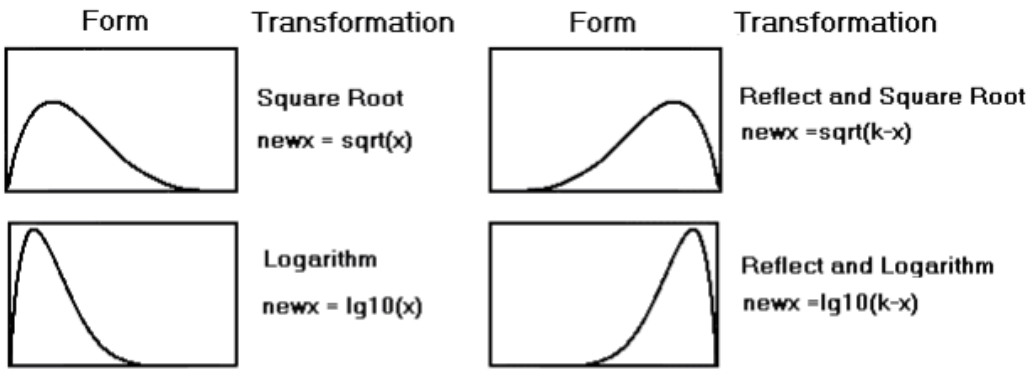


Figura 1: Representación de las distintas distribuciones posibles con la transformación más adecuada en cada caso para conseguir normalidad.

Cuando se ha optado por tomar el logaritmo neperiano de las observaciones de una variable, ha habido variables que en su estado original tomaban el valor 0, por lo que ha existido un problema a la hora realizar dicha transformación, pues el logaritmo nepriano de 0 es igual a $-\infty$. Para solucionarlo, se ha creado un indicador. Éste ha tomado el valor 0 si una observación es igual a 0 o el valor 1 si la observación ha sido distinta de 0. Una vez construido el indicador, se toma el logaritmo neperiano de las observaciones de la variables y, en los casos en los que la observación es igual a 0, en lugar de aplicar el logaritmo, se ha mantenido el valor 0. En cualquier caso, la variable resultante interacciona con el indicador [1]. El motivo por el cual se ha realizado esta maniobra es porque a la hora de crear el modelo de predicción, cuando las variables interaccionan con los indicadores, si una observación es igual a 0, se obtiene el siguiente modelo:

$$y = \beta_0, \tag{1}$$

y si una observación es distinta de 0 se obtiene el siguiente modelo:

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + ... + \beta_n \ln(x_n), \tag{2}$$

donde $\beta_0$ corresponde a la constante del modelo, $\beta_i$ corresponde a los parámetros de cada una de las variables $x_i$ y $\ln(x_i)$ corresponde con el logaritmo neperiano de cada una de las 69 variables.

## 3.1. Modelo de regresión lineal

Una vez transformadas todas las variables, se inicia una de las partes más importantes de este trabajo: la selección de variables del modelo de regresión lineal. Las variables que han formado el mejor modelo de regresión lineal han sido las variables que se han utilizado en el algorítmo *Gradient Boosting y en la* Redes Neuronal. Para la selección de variables, se ha utilizado un procedimiento contenido dentro del software estadístico SAS v9.2, llamado GLMSELECT [2]. Este procedimiento requiere de un número o semilla aleatoria, que inicializa el proceso, y devuelve un modelo formado por las varibles que en conjunto tienen un menor valor del Criterio de Informacion de Akaike ($AIC$) definido por la ecuación (3):

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2p, \tag{3}$$

donde $p$ es el número de parametros, $n$ el numero de observaciones y $SSE$, es el error de la suma de cuadrados, definido por la ecuación (4):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{4}$$

Por lo tanto, para obtener una lista con posibles modelos de regresión, y sabiendo que la semilla aleatoria que inicializa el proceso puede hacer variar

en gran medida los resultados, se prepara una colección de semillas aleatorias, obteniendo así un modelo de regresión para cada una de ellas.

De entre todos esos posibles modelos, se ha calculado la frecuencia de aparición, seleccionando sólo aquellos que hayan aparecido al menos dos veces y, mediante la técnica de validación cruzada que cuantifica el error cuadrático medio de un modelo de predicción, se ha seleccionado el modelo de regresión lineal que ha cometido un menor error cuadrático medio.

Hay que señalar que en todo momento, ya sea primero en la construcción del modelo de regresión lineal o posteriormente en las técnicas algorítmicas, se han utilizado tres particiones de los datos. Estas particiones son los datos de entrenamiento o *training*, los datos de validación y los datos de prueba. Los datos de entrenamiento suponen el 60 % del total de los datos y sirven para crear uno o varios posibles modelos en el caso de la regresión lineal o bien para entrenar los algorítmos en el caso del *Gradient Boosting* y las *Redes Neuronales*. Los datos de validación suponen el 20 % del total de los datos y sirven para ver cómo de bien funcionan los distintos modelos generados con datos que no haya visto nunca. Esta partición servirá para elegir, en función del error que cometan, el modelo más adecuado, en el caso de la regresión, el número de nodos ocultos en las *Redes Neuronales* y el número de hojas en el algoritmo *Gradient Boosting*. Por último, la partición de los datos de prueba, que ocupa el último 20 % de los datos, sirve para obtener una estimación más realista del error que se cometería en la predicción de la tasa de retorno de la carta del Censo de los Estados Unidos [1].

La técnica de validación cruzada previamente mencionada viene representada gráficamente en la Fig.2, y se basa en realizar una partición de los datos de validación, que son con los que se trabaja en esta parte, en $K$ subconjuntos, donde $K-1$ subconjuntos se utilizarán para entrenamiento y el subconjunto restante, para validación. Este proceso, conocido como $K-fold$, es repetido $K$ veces, utilizando cada uno de los subconjuntos una vez como submuestra de validación. A continuación se realiza la media aritmética de los $K$ resultados obtenidos durante el proceso para obtener una única estimación del error que produce el modelo de regresión en cuestión [3]. Además, al igual que con el procedimiento GLMSELECT, este proceso requiere un número o semilla aleatoria, de modo que si en lugar de introducir una única semilla, utilizamos un rango de $n$ semillas, para un mismo modelo, obtendremos $n$ estimaciones distintas del error. Éstas permiten visualizar graficamente, mediante un diagrama de cajas, una estimación de su varianza, su media y su mediana.

La técnica de validación cruzada ha sido utilizada a lo largo de todo el trabajo, pues sirve para cuantificar el error cuadrático medio de cualquier técnica predictiva, lo que permite poder comparar y ver cuál de ellas es la mejor.

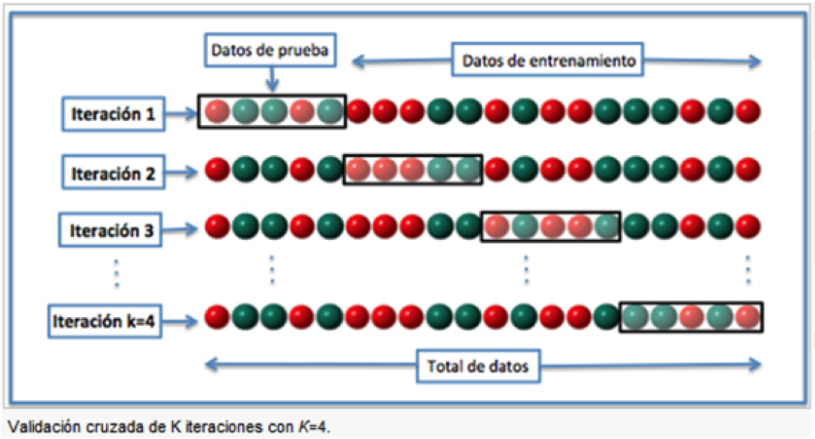


Figura 2: Técnica de validación cruzada *K-fold.*

Sin embargo, una inspección visual de los diagramas puede ser insuficiente a la hora de valorar qué modelo comete menor error, pues a simple vista podría parecer que existe una diferencia significativa entre las muestras de los errores de dos técnicas y, en realidad, no haberla. Para cuantificar objetivamente la decisión, se ha realizado un test para la media sobre muestras pareadas que permite saber si esas diferencias, que gráficamente pueden parecer significativas, realmente lo son.

Hay que señalar que el test es sobre muestras pareadas porque la técnica de validación cruzada, estableciendo un rango de $n$ semillas aleatorias, da lugar a valores como los que se muestran en el Cuadro 1. En ella se puede ver el error que comete el modelo o la técnica 1 con la semilla 1, el error que comete el modelo o la técnica 2 con la semilla 1, y así sucesivamente con cada semilla.

| | semilla 1 | semilla 2 | ... | semilla $n$ |
|---|---|---|---|---|
| Modelo/Técnica 1 | 20 | 27 | ... | 21 |
| Modelo/Técnica 2 | 22 | 21 | ... | 23 |

Cuadro 1: Datos ficticios que representa dos muestras pareadas.

Para saber si los valores medios de los modelos 1 y 2, mostrados en el Cuadro 1, son iguales, es necesario plantear un contraste de hipótesis bilateral. En él, según se muestra en la ecuación (5), se contrasta que los valores medios de los modelos 1 y 2 son iguales:

$$H_0 : \mu_1 = \mu_2, \tag{5}$$

frente a los que no lo son, tal y como se muestra en la ecuación (6):

$$H_1 : \mu_1 \neq \mu_2, \tag{6}$$

siendo $\mu_1$ la media de los valores del modelo 1 y $\mu_2$ la media de los valores del modelo 2.

El estadístico del contraste viene representado por la ecuación (7):

$$t = \frac{\bar{d}_i}{S_d/\sqrt{n}}, \tag{7}$$

donde $\bar{d}_i$ es la media de los $d_i$, que se calculan restando, para cada semilla, el valor que toma el modelo 1 menos el valor que toma el modelo 2. $S_d$ es la desviación típica de de los $d_i$, y $n$ es el número de semillas utilizadas en la técnica de validación cruzada.

Se rechazará $H_0$ si se cumple la deisgualdad $|t| > t_\alpha(n-1)$, para un nivel de significación $\alpha = 0{,}05$, donde $|t|$ se obtiene mediante la ecuación (7) , y $t_{0,05}(n-1)$ se obtiene de las tablas $\chi^2$. Si se cumple la desigualdad, se puede afirmar que, con un nivel de confianza del 95 %, existen diferencias significativas entre los valores medios de los modelos 1 y 2 mostrados en el Cuadro 1.

Por último, señalar que aunque este test se ha introducido en la descripción teórica de la parte perteneciente a los modelos de regresión lineal, dicho test se ha utilizado en cualquier parte del trabajo con el fin corroborar o desmentir las conclusiones que se puedan sacar de una comparación hecha a partir de una representación gráfica.

Una vez seleccionado el modelo de regresión lineal que comete menor error cuadrático medio, se procede a tratar de mejorar la predicción de dicho modelo. Para ello, se van a utilizar dos técnicas de predicción algorítmicas que a priori deberían cometer menor error cuadrático medio que un modelo de regresión lineal [4].

## 3.2. Algoritmo *Gradient Boosting*

La primera técnica es el algoritmo *Gradient Boosting*. El *Boosting* es una de las ideas más interesantes introducidas a lo largo de las últimas dos décadas [4]. En un principio, estos métodos fueron diseñados para problemas de clasificación, pero también pueden ser utilizados en problemas de regresión. El origen del *Boosting* viene de la combinación de clasificadores (o predictores si se trata de un problema de regresión) que en sí mismos no son muy potentes para conseguir un clasificador (o predictor) más potente.

Para la implementación de este algoritmo hay que disponer de un conjunto de datos de entrenamiento $(x_i, y_i)$ y es necesario establecer una función de coste para comparar el valor real con el valor estimado. Esta función de coste viene definida por la ecuación (8):

$$\frac{1}{2}\sum |y_i - f(x_i)|^2, \tag{8}$$

- **Estimación inicial:** Para que de comienzo el algoritmo, es necesario introducir una estimación de la tasa de retorno. Este paso se visualiza en la

ecuación (9), que muestra que la estimación inicial de la tasa de retorno es la media de las tasas de retorno.

$$\hat{y}_i^0 = \bar{y}, \tag{9}$$

- **Cálculo de los residuos:** El cálculo de los residuos se muestra en la ecuación (10):

$$r_i^{(m)} = y_i - \hat{y}_i^{(m)}, \tag{10}$$

  donde los residuos son el gradiente, dada la función de error vista en la ecuación (8).

- **Ajuste de los residuos:** Una vez calculados los residuos, hay que ajustarlos mediante árboles de regresión.

- **Actualización de las predicciones:** Tras ajustar los residuos, se actualizan las predicciones $\hat{y}_i$ utilizando la ecuación (11):

$$\hat{y}_i^{(m+1)} = \hat{y}_i + v \cdot r_i^{(m+1)}, \tag{11}$$

- **Vuelta al cálculo de residuos:** Una vez actualizadas las predicciones, se vuelve de nuevo a calcular los residuos, que van disminuyendo hasta alcanzar la convergencia.

## 3.3. *Redes Neuronales*

La segunda técnica es una *Red Neuronal artificial*, que fue desarrollada en dos campos diferentes - la estadística y la inteligencia artificial -, aunque basadas en modelos similares [4].

La idea principal de las redes neuronales es obtener combinaciones lineales de las variables independientes del modelo y utilizarlas como variables independientes derivadas para modelizar la variable dependiente como una función no lineal de esas variables independientes derivadas.

Para entender el funcionamiento de una *Red Neuronal*, es necesario observar la imagen mostrada en la Fig. 3. En ella, se puede ver el funcionamiento de una neurona cerebral, donde la información es recibida a través de las dendritas y procesada en el núcleo de la neurona, y la respuesta es emitida a través de los axones.

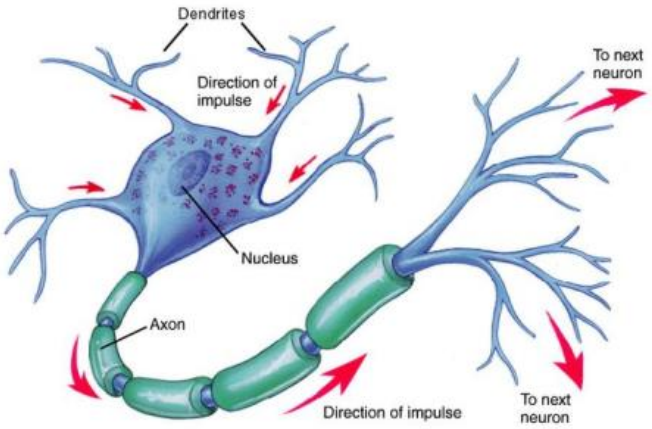


Figura 3: Diagrama de una neurona.

Este funcionamiento es el que se trata de imitar a través de la *Red Neuronal artificial*, y a través de la Fig. 4 se puede ver un esquema de su funcionamiento. En ella, se observa una primera capa donde se introduce la información, que simularía ser la dendrita en la neurona. A continuación, existe una capa oculta, que se correspondería con el núcleo de la neurona con nodos, que son el resultado de las combinaciones no lineales de las variables de la primera capa, y que servirán para obtener la última capa a través de la cuál se obtiene la respuesta. Esta última capa simularía ser el axón en la neurona de la Fig. 3.

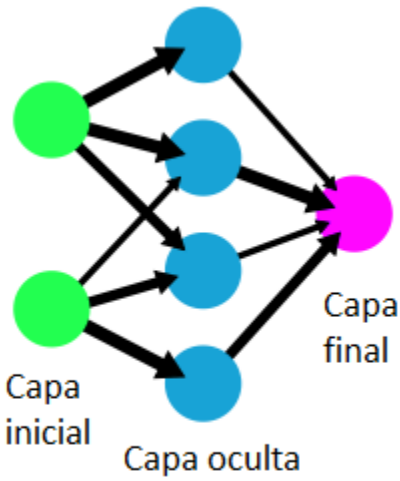


Figura 4: Diagrama de una *Red Neuronal artificial*.

Para implementar una *Red Neuronal artificial*, hay que establecer los siguientes parámetros:

$$\{\alpha_{0m}, \alpha_m; m = 1, 2, 3, ..., M\}\, M(p+1) ponderaciones, \tag{12}$$

$$\{\beta_{0m}, \beta_m; m = 1, 2, 3, ..., K\}\, K(M+1) ponderaciones, \tag{13}$$

Al estar trabajando con un problema de regresión, hay que establecer una función de coste; en este caso, la suma de cuadrados de los errores:

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2. \tag{14}$$

Una vez establecdia la función de coste, hay que derivar con respecto a $\alpha$ y con respecto a $\beta$ para realizar una minimización de dicha función.

$$\frac{\partial R_i}{\partial \beta_{km}} = -2(y_{ik} - f_k(x_i)) g_k(\beta_k^T z_i) z_{mi} = \delta_{ki} z_{mi}, \tag{15}$$

donde $z_{mi} = \sigma(\alpha_{0m} + \alpha_m^T x_i)$.

$$\frac{\partial R_i}{\partial \alpha_{ml}} = -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i)) g_k(\beta_k^T z_i) \beta_{km} \sigma(\alpha_m^T x_i) x_{il} = s_{mi} x_{il}, \tag{16}$$

donde $s_{mi} = \sigma(\alpha_m^T x_i) \sum_{k=1}^{K} + \beta_{km} \delta ki$.

Una vez establecidos los parámetros, la función de coste, y tras haber derivado con respecto a $\alpha$ y con respecto a $\beta$, se comienza la optimización de la función de coste. Para ello, es necesario entrar en un proceso iterativo en el cual se inicializan las ponderaciones de los parámetros, se calcula el error cuadrático medio que se comete, y se actualizan los parámetros gracias a las derivadas parciales llevadas a cabo. De este modo, el algoritmo, poco a poco comete menor error cuadrático medio hasta la convergencia.

Señalar que en este proyecto se van a trabajar dos técnicas de optimización; el algoritmo de Levenberg-Mardquardt y la técnica *Backpropagation*.

La propagación hacia atrás o *Backpropagation*, es una técnica en la que una vez que se ha introducido la información, ésta se propaga desde la primera capa a través de las capas superiores de la red, hasta generar una salida [4]. El valor de salida se compara con el valor real y se calcula el error para cada una de las salidas. Las salidas de error se propagan hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida. Sin embargo, las neuronas de la capa oculta sólo reciben una parte del error, basándose aproximadamente en la contribución relativa que haya aportado cada neurona a la salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido el error que describa su contribución relativa al error total. La importancia de este proceso consiste en que, a medida que se entrena la red, las neuronas de las capas intermedias se organizan a sí mismas de

tal modo que las distintas neuronas aprenden a reconocer distintas características del espacio total de entrada.

En cuanto al algoritmo de Levenberg-Mardquardt, es una técnica iterativa que localiza el mínimo de una función multidimensional que viene expresada por la suma de cuadrados de los errores. Su descripción es muy similar a la dada para describir en términos generales una *Red Neuronal*, pues se ha convertido en una de las técnicas de optimización más utilizadas [6].

## 3.4. Sobreajuste de un modelo de predicción

Una vez introducidos los fundamentos teóricos del modelo de regresión lineal, el algoritmo *Gradient Boosting* y la *Red Neuronal* que se han utilizado en este trabajo cabe preguntarse: Si se introducen más parámetros en el modelo de regresión lineal, más hojas en el algoritmo *Gradient Boosting* y más nodos ocultos en la *Red Neuronal*, ¿no disminuirá el error cuadrático medio?

La respuesta a esta pregunta es sí. Introducir más parámetros, más hojas o más nodos podría disminuir el error cuadrático medio; sin embargo, el modelo o los algoritmos estarían sobreajustados. Es por ello por lo que uno de los puntos clave de este trabajo consiste en estimar correctamente el número adecuado de parámetros, nodos ocultos u hojas, ya que así se estaría evitando el sobreajuste del modelo. Esto se debe a que si permitimos al algoritmo o a la red entrenarse demasiado, o al modelo le introducimos demasiados parámetros, éstos pueden quedar ajustados a unas características muy concretas de los datos, y a la hora de realizar predicciones con datos nuevos, el error iría en aumento. Esta idea queda explicada de forma visual en la Fig. 5, donde el ajuste A representa un ajuste mucho más simple que el ajuste B. Se dice que un modelo está sobreajustado cuando se ajusta perfectamente a los datos con los que se ha entrenado. Esta situación, que claramente representa el ajuste B, comete un error mínimo con datos de entrenamiento. Sin embargo, si se utilizan datos de validación, que el modelo no ha visto nunca, y el modelo se encuentra ajustado a una característica muy particular de los datos de entrenamiento que no se encuentra en los datos de validación, el error se dispararía.

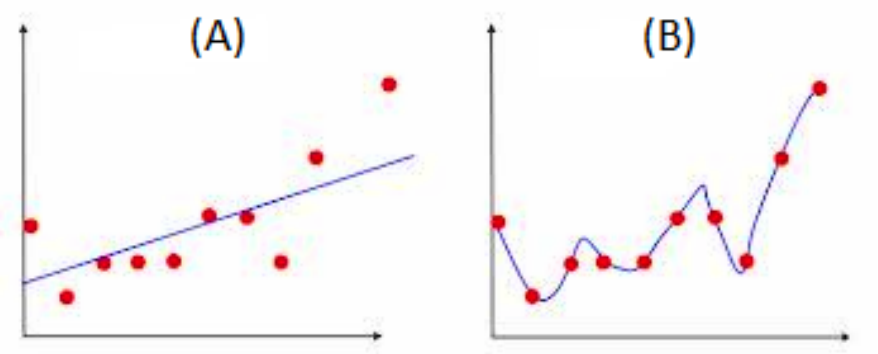


Figura 5: (A) Gráfica con un modelo sin sobreajuste. (B) Gráfica con un modelo con sobreajuste.

Una de las técnicas más utilizadas para detectar si un modelo está o no sobreajustado, consiste en representar gráficamente el error que comete el modelo a medida que aumenta el número de parámetros que lo componen, y en función del conjunto de datos con el que se trabaje. En Fig. 6 se puede ver cómo el error que comete el modelo utilizando los datos de entrenamiento, representado por la línea azul, siempre disminuye, ya que a medida que aumenta el número de parámetros, el modelo se va ajustando cada vez mejor a los datos. En cuanto al error cometido en datos de validación, representado por la línea roja, se puede ver como hay un punto a partir del cual el error, en lugar de disminuir, comienza a aumentar. Esto se debe a que, a partir de ese punto, el modelo empieza a estar demasiado bien ajustado a los datos, y es en ese punto cuando pueden no aparecer observaciones que sí habían aparecido en los datos de entrenemiento y que causarían un aumento del error [4].

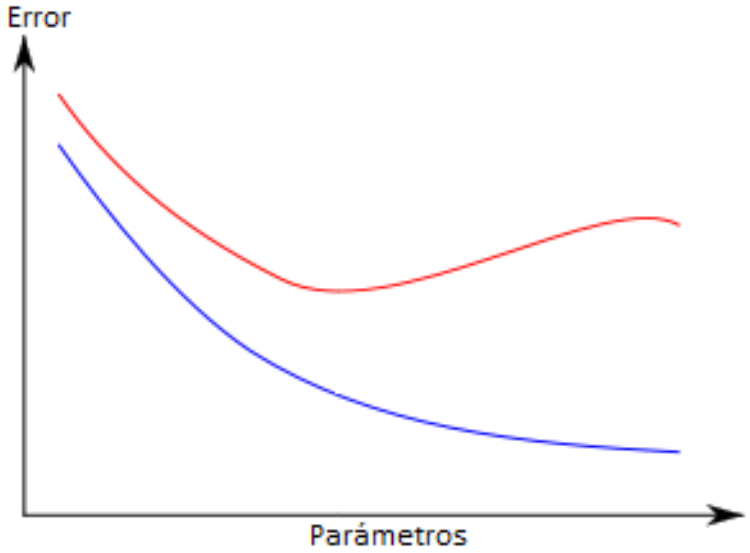


Figura 6: Comportamiento del error, representado en el eje de abcisas. en funciôm del número de parámetros, representado en el eje de ordenadas, y según el conjunto de datos con el que se trabaje.

## 3.5. Comparación de las técnicas de predicción

Finalmente, utilizando el procedimiento GMAP [2], se han realizado 2 tipos de representaciones gráficas sobre el mapa de los Estados Unidos. El primer tipo de representaciones gráficas complementa la comparación final entre el error cuadrático medio que cometen las distintas técnicas de predicción utilizadas, pues se ha realizado una representación del error que se comete por hogar en cada estado del país con cada técnica utilizada.

El segundo tipo de representación gráfica presenta la mediana de la tasa de retorno de la carta del censo de los Estados Unidos, por familia y por estado, primero utilizando las tasas de retorno reales, y luego utilizando las tasas de retorno estimadas en función de la técnica utilizada. De esta forma se puede ver como se asemajan las tasas de retorno estimadas a las tasas de retorno reales.

# 4. Resultados

Tras haber presentado los fundamentos teóricos necesarios para realizar este trabajo, se procede explicar y mostrar los resultados obtenidos.

## 4.1. Imputación de valores perdidos o *missing*

En primer lugar, hacemos una depuración de los datos dada la existencia de un elevado número de valores perdidos o *missing*. Recordar que la imputación de los valores *missing* se ha realizado utilizando estimaciones censales propocionadas por la oficina del Censo de los Estados Unidos.

## 4.2. Transformación de variables

Una vez imputados los valores *missing*, se muestra, a través de la Fig, 7, una visualización gráfica de la distribución de algunas de las variables independientes. En ellas se ve que existe una gran diferencia entre las escalas de las variables, lo que dificulta ver su distribución de cada variable en los histogramas. Sin embargo, en aquellas en las que se puede ver la distribución se puede apreciar una asimetría positiva, lo que, según la Fig. 1, se podría solucionar tomando logaritmos de los valores de las variables, tomando la raíz cuadrada o utilizando la transformación inversa. En este caso, se ha optado por transformar logaritmicamente los valores de todas las variables independientes, ya que no sólo se consigue normalizar la distribución de la variable, sino que además, datos medidos en distintas escalas pasan a ser comparables entre sí y el efecto de posibles valores extremos o atípicos queda reducido.

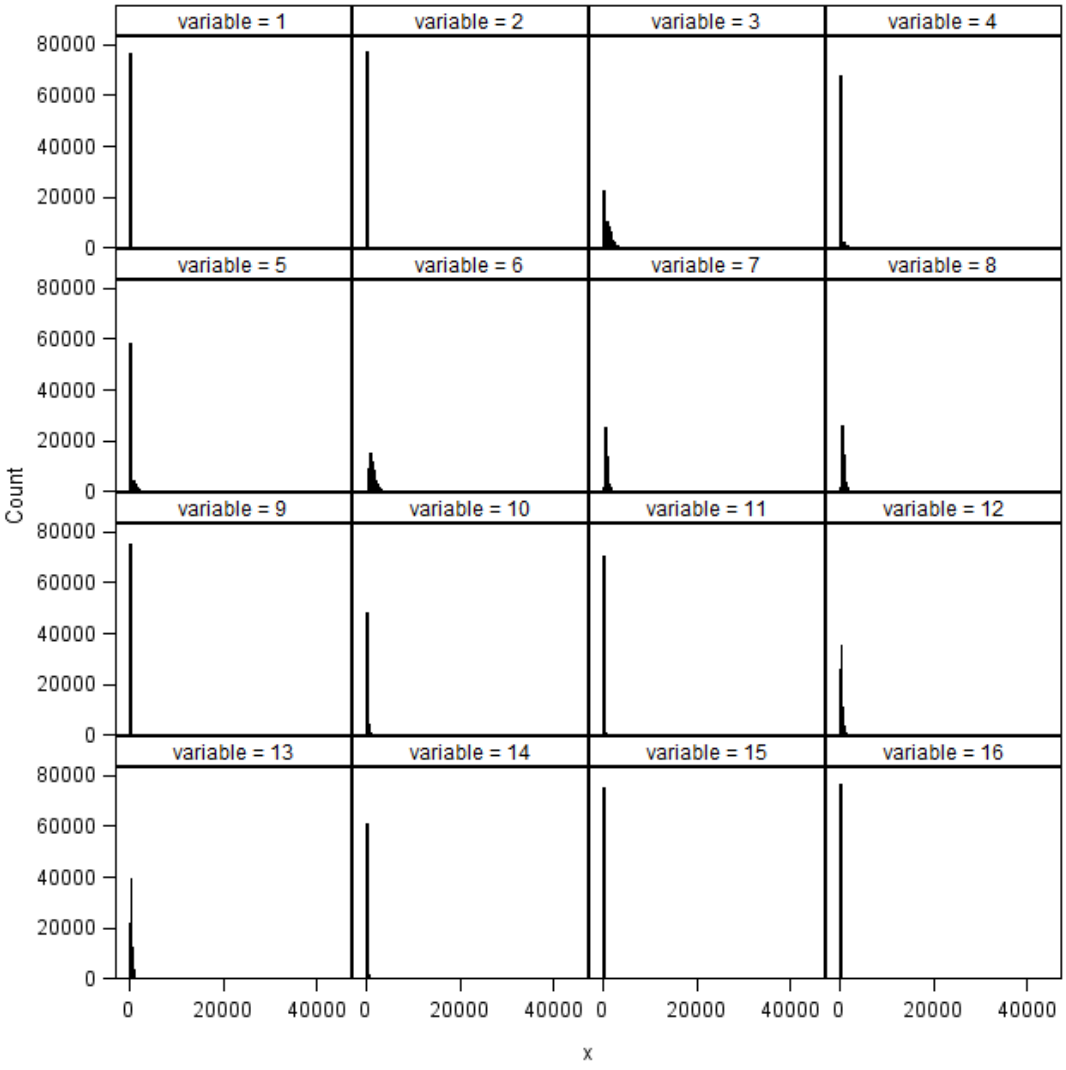


Figura 7: Diagrama de dispersión de las variables independientes $\ln(x_1) - \ln(x_{16})$.

A continuación, en la Fig. 8, se muestran los histogramas algunas de las variables independientes una vez se han transformado logaritmicamente sus valores. En este caso sólo se muestran aquellas observaciones cuyo indicador es igual a uno, lo que implica que la variable $x_2$ no ha sido representada pues todas sus observaciones son igual a 0. El motivo por el cual no se muestran los valores iguales a 0 es porque en los datos, si el indicador creado es igual a 0, el modelo de predicción en ese caso estará compuesto tan sólo por una constante, tal y como se observa en la ecuación (1), mientras que si el indicador es igual a uno, el modelo de predicción estará compuesto por la constante y todas las variables que lo conformen, tal y como se observa en la ecuación (2). Además, en este caso, lo que realmente interesaba era ver si se había conseguido o no normalizar la distribución de los valores con indicador igual a 1, y la inclusión de aquellos valores con indicador igual a 0 hubiese dificultado la visualización del histograma, tal y como sucede en la Fig. 7.

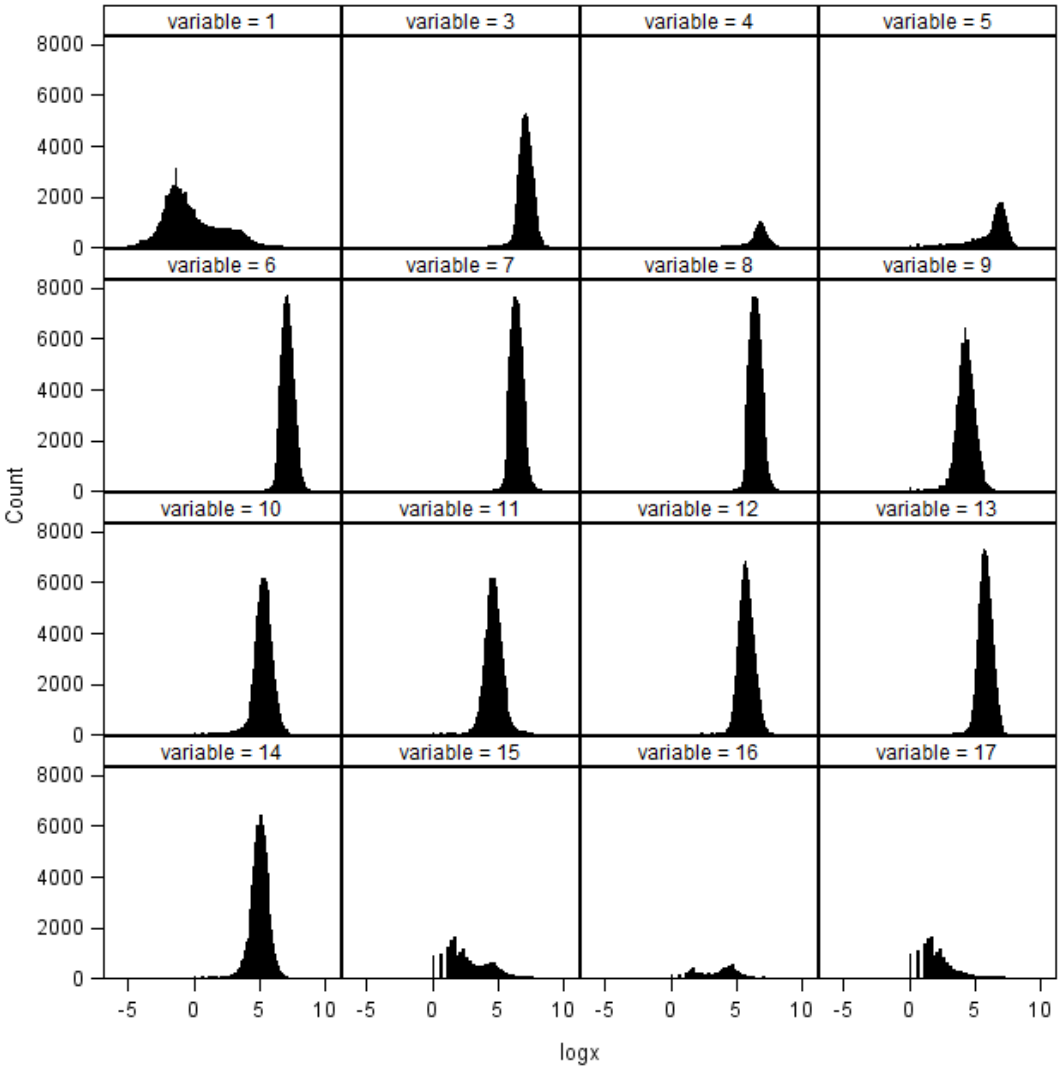


Figura 8: Diagrama de dispersión de las variables independientes $\ln(x_1) - \ln(x_{16})$.

A la vista de los resultados, se puede concluir que se ha conseguido normalizar la distribución de las observaciones con indicador igual a uno en la mayor parte de las variables.

Las Fig. 9 y 10 muestran, mediante diagramas de dispersión, la relación de 4 variables con la vaiable dependiente. Fig. 9 contiene las variables en su estado original, mientras que la Fig. 10 contiene las variables y tras haber tomado el logaritmo neperiano de sus valores. En ellas se puede ver que la normalización de dichas variables, a través del logaritmo neperiano de sus valores, genera una relación más lineal con la variable dependiente.

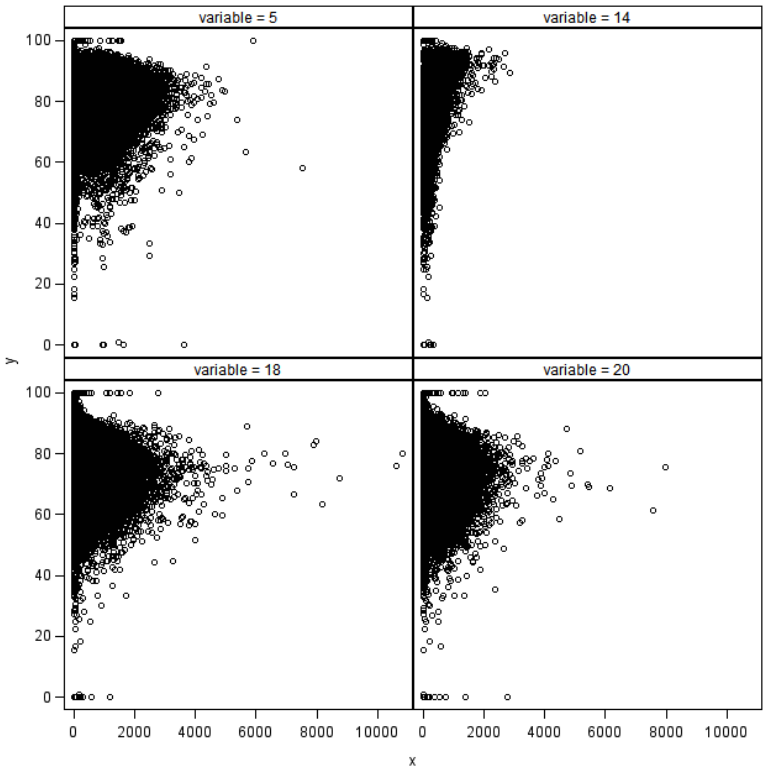


Figura 9: Diagrama de dispersión de las variables $x_5, x_{14}, x_{18}$ y $x_{20}$.

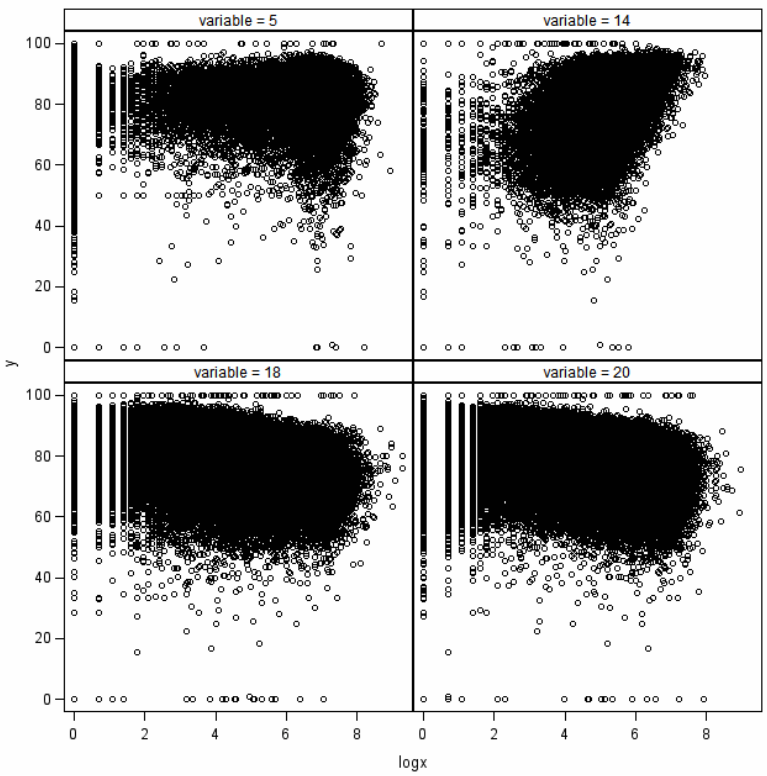


Figura 10: Diagrama de dispersión de las variables $\ln(x_5), \ln(x_{14}), \ln(x_{18})$ y $\ln(x_{20})$.

En cuanto a la variable dependiente, posee asimetría negativa en su distribución. Tomando como referencia la Fig. 1, y con el fin de conseguir una distribución más normalizada, existe la posibilidad de tomar logaritmos neperianos o bien la raíz cuadrada de los valores de la variable.

Dado que tanto la transformación logarítmica como la raíz cuadrada parecen ser una buena opción, se han utilizado ambas transformaciones. Recordar que en ambos casos, al tratar de transformar una variable con asimetría negativa, la transformación no se realiza directamente sobre $y$ sino sobre $K - y$, siendo $K$ en valor máximo de dicha variable. Dichas transformaciones sobre la variable dependiente se muestran a través de ecuaciones (17) y (18):

$$y2 = \sqrt{100 - y}, \tag{17}$$

$$y3 = \ln(100 - y), \tag{18}$$

donde $y$ es la variable dependiente sobre la que se aplican las transformaciones.

Una vez transforma la variable independiente, se puede concluir que, con ambas transformaciones, se obtiene una distribución más normalizada que la distribución original.

## 4.3. Búsqueda del modelo de regresión lineal

A continuación, se realiza la primera búsqueda de un modelo de regresión lineal. Para ello, se ha utilizado como variable dependiente la variable transformada $y2$, definida en la ecuación (17). En cuanto a las variables independientes, se han utilizado las variables transformadas $\ln(x_1) - \ln(x_{69})$, que hay que recordar que interaccionan con los indicadores $indi_1 - indi_{69}$.

Al introducir toda esta información en el procedimiento GLMSELECT, y estableciendo un rango de 300 semillas aleatorias, se generan 300 posibles modelos. Tras haber ordenado por frecuencia de aparición y habiendo seleccionado sólo aquellos que aparecen al menos dos veces, mediante validación cruzada, se lleva a cabo una comparación entre todos los modelos. Los 11 modelos que presentan un menor error cuadrático medio se muestran en la Fig. 11, en la que se observa que los modelos 31 y 53 son los que cometen el menor error cuadrático medio.

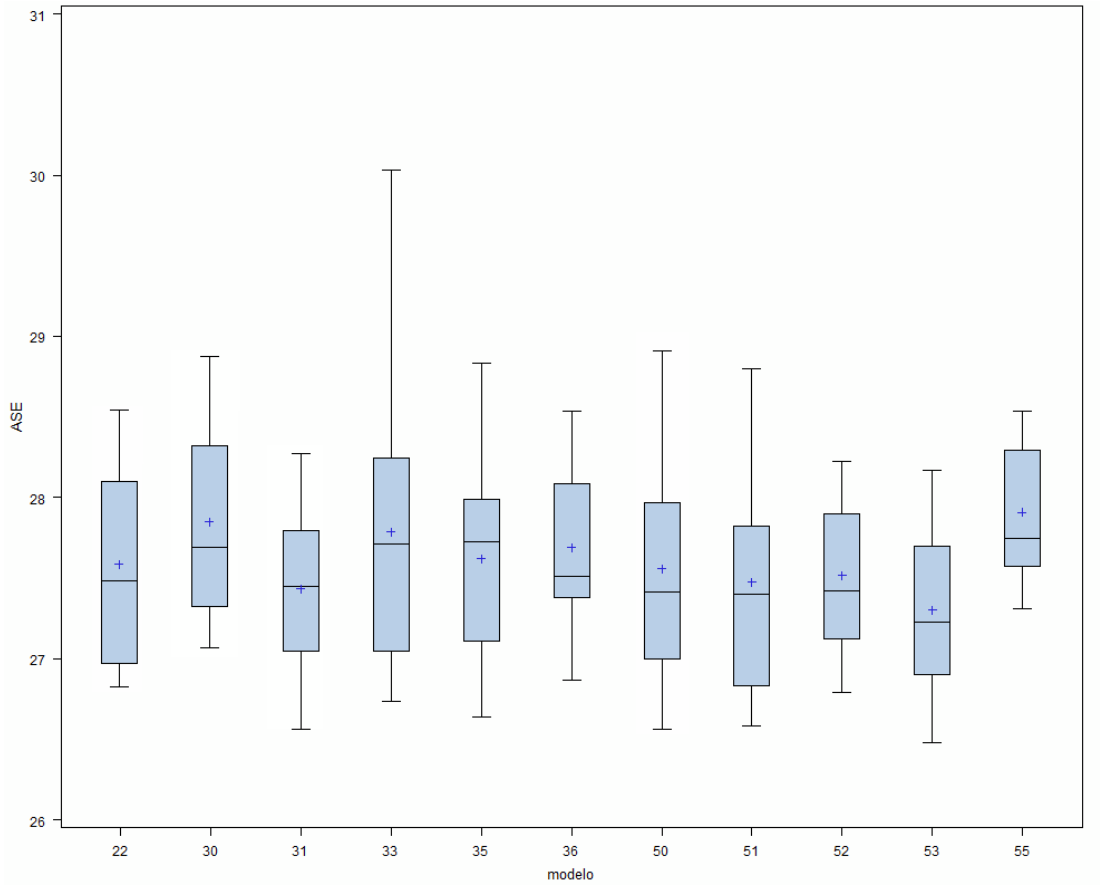


Figura 11: Gráfico con los 11 mejores modelos, comparados mediante validación cruzada, de la primera búsqueda

Para determinar qué modelo es el adecuado, se realizó un test para la media sobre muestras pareadas, obteniendo los resultados expuestos en el Cuadro 2.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{31}$=$\mu_{53}$ | -4.75 | 0.008 |

Cuadro 2: Resultados del test para la media sobre muestras pareadas de los modelos 31 y 53, en la que se establecen diferencias significativas entre los errores cuadrático medios de ambos modelos

Con un *p-valor* igual a 0.08, se puede rechazar la hipótesis nula en la que $\mu_{31}$=$\mu_{53}$, lo que permite conluir que existen diferencias significativas entre los modelos 31 y 53; eligiendo este último como el modelo óptimo en esta primera búsqueda.

Para la segunda búsqueda, se han utilizado exclusivamente las variables originales, tanto la dependiente como las independientes. El objetivo de esta búsqueda ha sido ver si realmente las transformaciones realizadas sobre la variable dependiente y sobre las variables independientes han supuesto una mejora en la capacidad predictiva de un modelo de regresión lineal, con variables transformadas, frente a otro modelo de regresión lineal con variables sin modificar.

Al igual que en la primera búsqueda, se han introducido las variables en el procedimiento GLMSELECT. Establecemos el mismo rango de 300 semillas aleatorias, se han obtenido 300 posibles modelos de regresión lineal. De nuevo, ordenando por frecuencia de aparición, y seleccionando aquellos modelos que aparecían al menos 2 veces, se ha utilizado la técnica de validación cruzada para cuantificar el error en la predicción que ha cometido cada modelo.

De los 300 posibles modelos de regresión lineal, se ha realizado una comparación entre todos ellos eligiendo, en este caso, los siete modelos que presentaban un menor error cuadrático medio. Esta comparación se muestra en la Fig. 12, donde se ve que los modelos 50 y 51 son lo que presentan un menor error cuadrático medio.

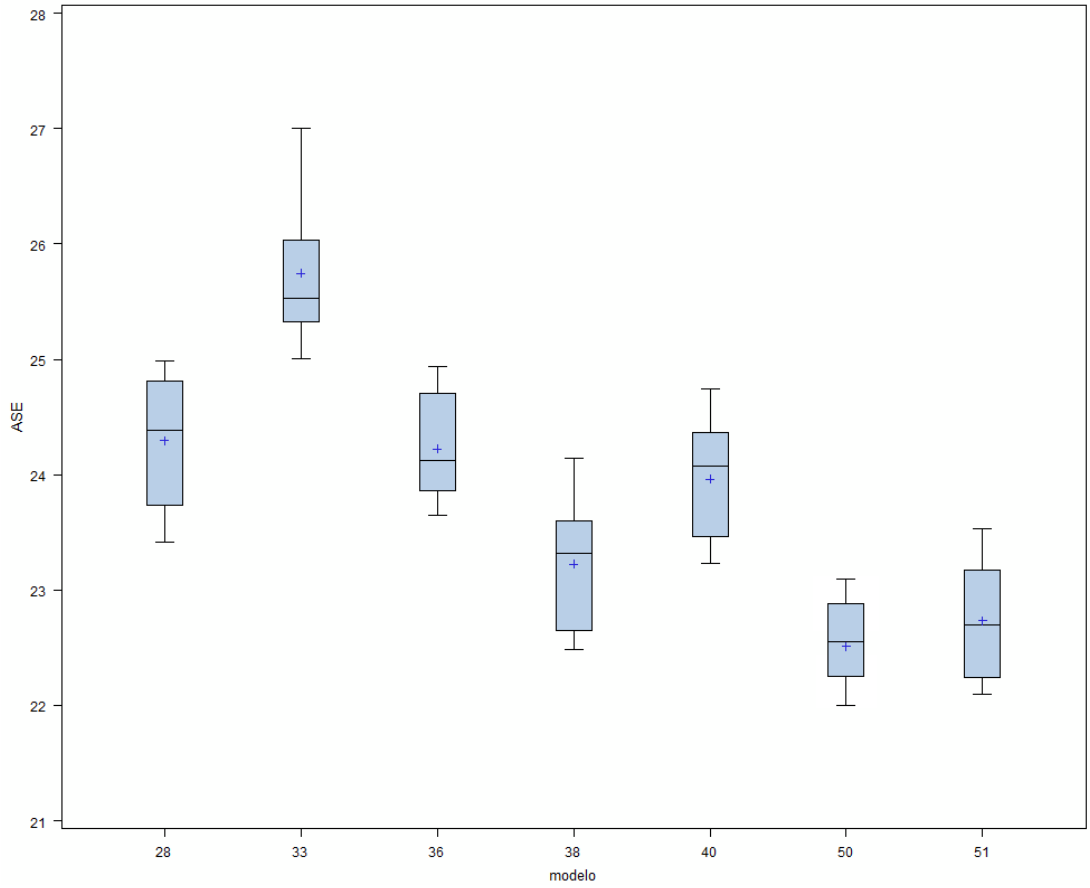


Figura 12: Gráfico con los 7 mejores modelos, comparados mediante validación cruzada, de la segunda búsqueda

Para determinar qué modelo es el adecuado, se ha realizado un test para la media sobre muestras pareadas, obteniendo los resultados que se muestran en el Cuadro 3. En ellos se observa que el *p-valor* del contraste ha sido igual a 0.8427, lo que no ha permitido establecer diferencias estadísticamente significativas entre los modelos 50 y 51. Sin embargo, gráficamente se puede apreciar que el modelo 50 ha tenido menor varianza, lo que lo ha convertido en el modelo óptimo en esta segunda búsqueda.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{50}$=$\mu_{51}$ | 0.2 | 0.8427 |

Cuadro 3: Resultados del test para la media sobre muestras pareadas de los modelos 50 y 51, en la que no se establecen diferencias significativas entre los errores cuadrático medios de ambos modelos

Para concluir la búsqueda de un modelo de regresión lineal se ha realizado una tercera búsqueda de este tipo de modelos de regresión lineal. Para ello, se ha utilizado como variable dependiente la variable transformada $y3$, explicada en la ecuación 18. En cuanto a las variables independientes, se han utilizado de nuevo las variables transformadas $\ln(x_1) - \ln(x_{69})$. El objetivo de esta búsqueda es ver si la transformación llevada a cabo sobre la variable dependiente, explicada en la ecuación (18), tiene mejores resultados que la transformación explicada en la ecuación (17).

Tras introducir todas las variables en el procedimiento GLMSELECT, y estableciendo el mismo rango de 300 semillas aleatorias, se han obtenido 300 posibles modelos de regresión. Tras ordenar los modelos de regresión por frecuencia de aparición y seleccionando aquellos modelos con una frecuencia igual o superior a 2, se ha recurrido de nuevo a la técnica de validación cruzada, ya que permite cuantificar el error que comete cada modelo de regresión. Los resultados se pueden ver en la Fig. 13. En ellos se aprecia que los modelos 39 y 60 son lo que presentan un menor error cuadrático medio.

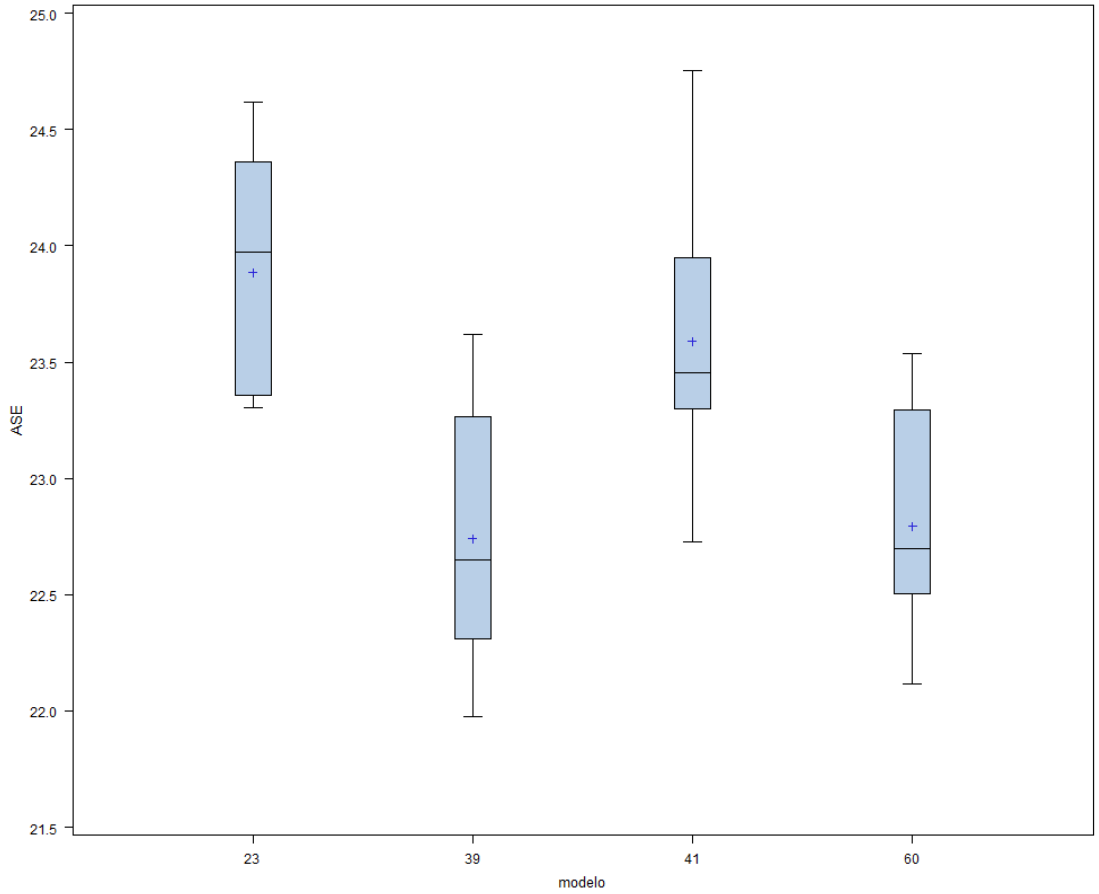


Figura 13: Gráfico con los 4 mejores modelos, comparados mediante validación cruzada, de la tercera búsqueda

Para determinar qué modelo es el adecuado, se ha realizado un test para la media sobre muestras pareadas, obteniendo los resultados que se muestran en el Cuadro 4. En ellos se ve que el *p-valor* del contraste es igual a 0.7579, lo que ha impedido establecer diferencias estadísticamente significativas entre los modelos 39 y 60. Sin embargo, gráficamente se aprecia que el modelo 60 tiene menor varianza, lo que lo ha convertido en el modelo óptimo en esta tercera búsqueda.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{39}$=$\mu_{60}$ | 0.32 | 0.7579 |

Cuadro 4: Resultados del test para la media sobre muestras pareadas de los modelos 39 y 60, en la que no se establecen diferencias significativas entre los errores cuadrático medios de ambos modelos

Una vez obtenidos 3 modelos con características distintas, uno por cada búsqueda realizada, se ha hecho una comparación utilizando la técnica de validación cruzada para saber qué modelo es el que comete menor error cuadrático medio.

Esta comparación se muestra en la Fig. 14, y se ve que los modelos de la segunda y tercera búsqueda son los que cometen menor error cuadrático medio.

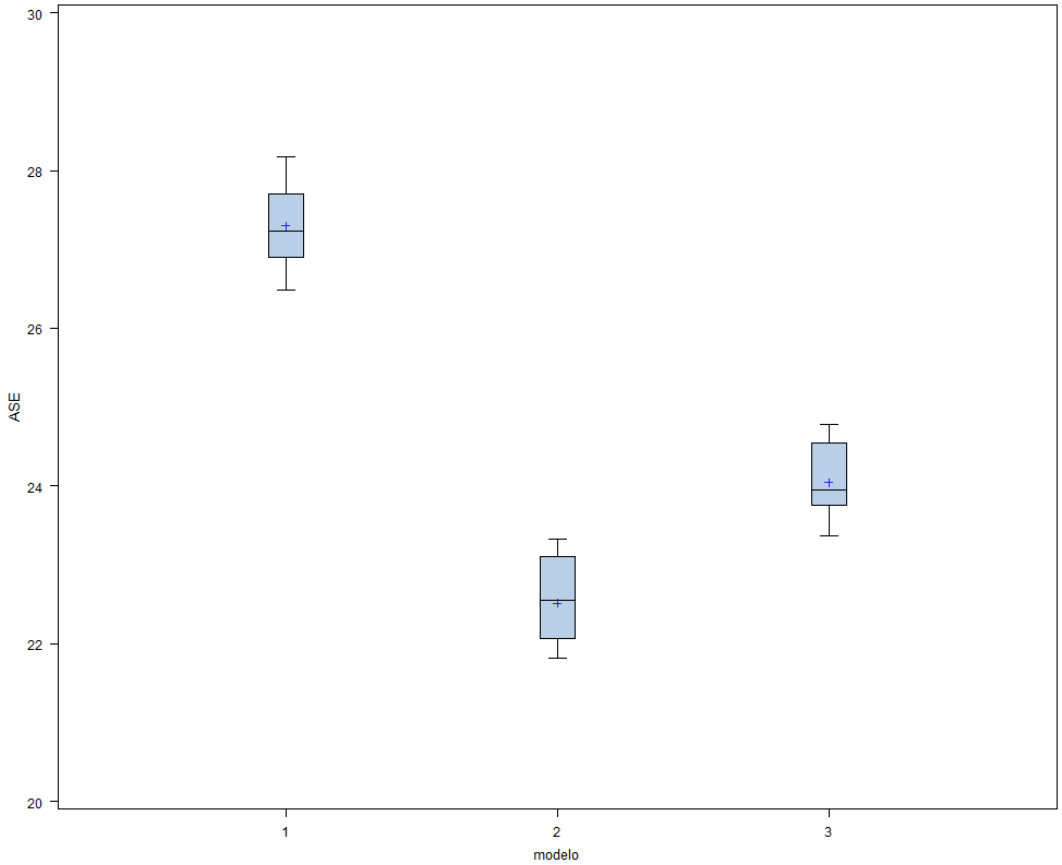


Figura 14: Gráfico con el mejor modelo de cada una de las búsquedas.

Gráficamente se aprecia que el modelo de la búsqueda 2, definido como "Modelo 2", comete un menor error que el modelo de la búsqueda 3, definido como "Modelo 3". Para saber si las diferencias son estadísticamente significativas, se ha realizado un test para la media sobre muestras pareadas, cuyos resultados se muestran en el Cuadro. 5.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_2$=$\mu_3$ | 6.81 | $< .0001$ |

Cuadro 5: Resultados del test para la media sobre muestras pareadas de los modelos 2 y 3, en la que no se establecen diferencias significativas entre los errores cuadrático medios de ambos modelos.

El *p-valor* obtenido, inferior a 0.05, ha permitido establecer diferencias estadísticamente significativas entre ambos modelos, lo que significa que el modelo de regresión lineal que comete menor error cuadrático medio es el que se ha obtenido en la segunda búsqueda. Éste modelo queda definido a través de la ecuación (19).

$$y = \beta_0 + \beta_1 \times (x5 \times f2) + \beta_2 \times (x14 \times f2) + \beta_3 \times (x18 \times f2) + \beta_4 \times (x20 \times f2) + \beta_5 \times (x40 \times f2) + \beta_6 \times (x43 \times f2) + \beta_7 \times (x46 \times f2) + \beta_8 \times (x50 \times f2) + \beta_9 \times (x56 \times f2) + \beta_{10} \times (x68 \times f2) + \varepsilon_i, \quad (19)$$

donde $y$ representa la variable dependiete, $\beta_i$ representa los parámetros de la ecuación, *f2* representa el estado, *x5* representa el condado, *x14* representa el total de la población censada en 2010 en cada hogar, *x18* representa el número total de hombres en el hogar, *x20* representa el número total de mujeres en el hogar, *x40* representa las personas de 65 años o más en el hogar, *x43* representa el origen hispano, *x46* representa el total de la población hispana en el hogar, *x50* representa el número de afroamericanos en el hogar, *x56* representa el número de asiáticos en el hogar y *x68* representa el úmero de personas que hablan otro idioma a parte de inglés en el hogar.

## 4.4. Implementación de algoritmo *Gradient Boosting*

Una vez hecha la búsqueda de un modelo de regresión lineal, se han utilizado otras técnicas con el fin de ver si éstas ofrecen una mejor capacidad predictiva.

En primer lugar, se han introducido las variables que componen el mejor modelo de regresión en el algoritmo *Gradient Boosting*, que presenta dos parámetros desconocidos que hay que estimar. Estos parámetros son el número de hojas que componen los árboles de decisión con los que funciona el algoritmo, tal y como se ha explicado en la sección 3, y la constante de regularización, que hace frente al sobreajuste del modelo.

A continuación, se muestra la Fig. 15, donde, tras haber establecido una constante de regularización $v$=0.01, se visualiza el error cuadrático medio que se comete en función del número de hojas que contengan los árboles de decisión para datos de entrenamiento.

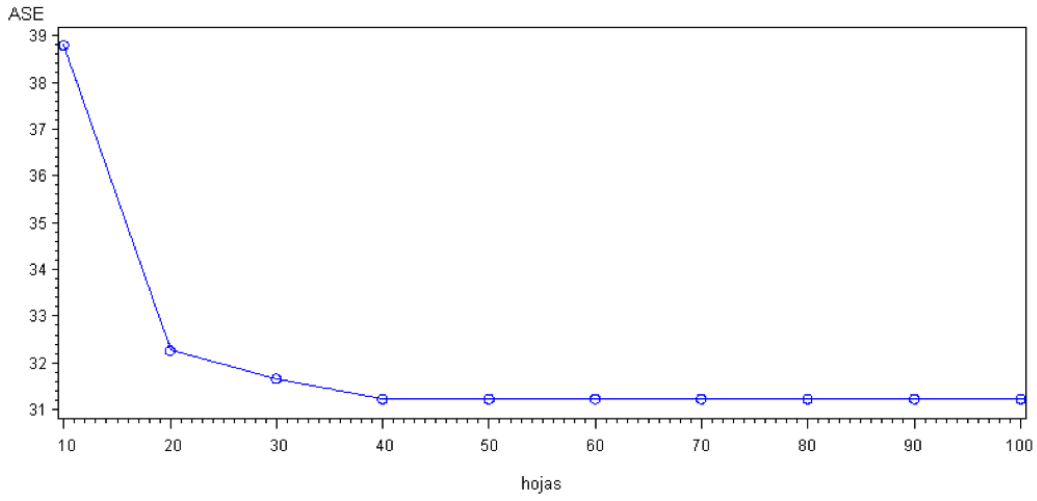


Figura 15: Representación del error que se comete según el número de hojas que se utilicen en los árboles de decisión y con una constante de regularización $v$=0.01 con datos de entrenamiento.

Se observa que, para datos de entrenamiento, la estabilización del error cuadrático medio se alcanza en 40 hojas. Si en lugar de utilizar una constante de regularización $v$=0.01, utilizamos una constante de regularización $v$=0.1, se obtienen los resultados que se muestran en la Fig. 16. En ella se observa que el error cuadrático medio se estabiliza también en 40 hojas.

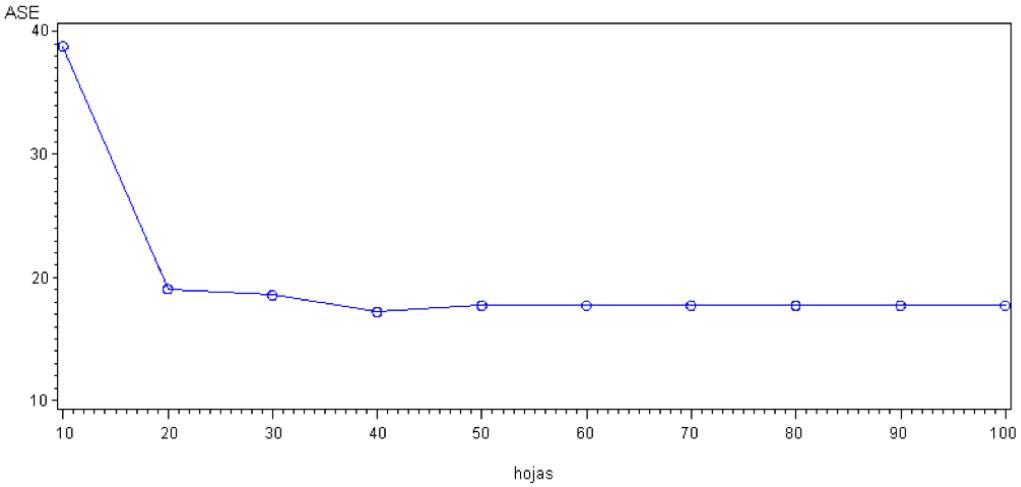


Figura 16: Representación del error que se comete según el número de hojas que se utilicen en los árboles de decisión y con una constante de regularización $v$=0.1 con datos de entrenamiento.

Si en lugar de utilizar datos de entrenamiento, utilizamos datos de validación y ambas constantes de regularización, se obtienen los resultados que se muestran en las Fig. 17 y 18. En ellas se observa cómo, para ambas constantes de regularización, el error cuadrático medio desciende conforme aumenta el número de hojas, hasta que se alcanzan las 40 hojas cuando el error cuadrático medio comienza a aumentar.

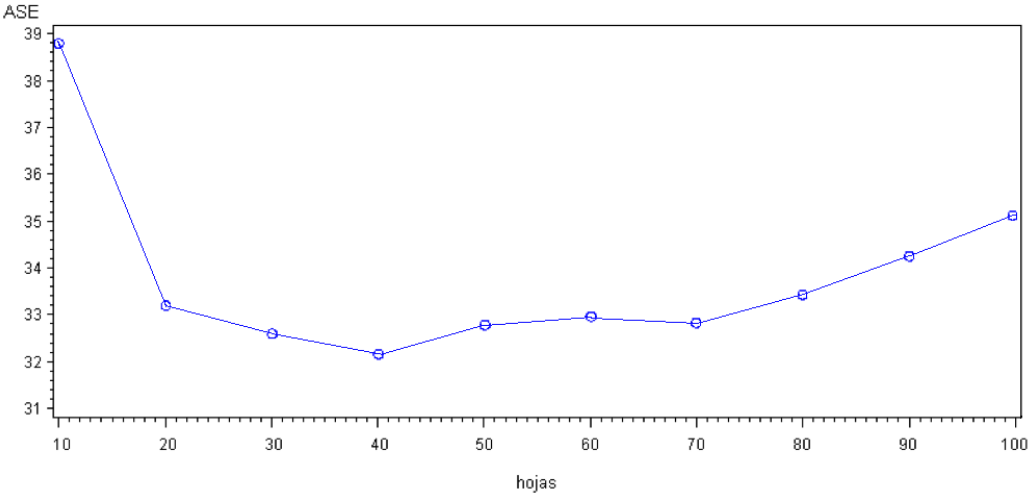


Figura 17: Representación del error que se comete según el número de hojas que se utilicen en los árboles de decisión y con una constante de regularización $v$=0.01 con datos de validación.

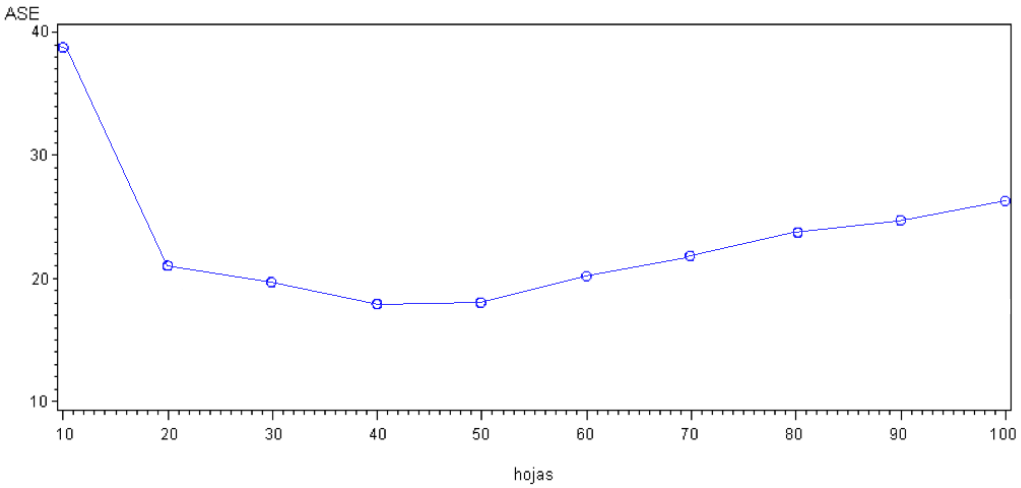


Figura 18: Representación del error que se comete según el número de hojas que se utilicen en los árboles de decisión y con una constante de regularización $v$=0.1 con datos de validación.

Para saber qué constante de regularización es la adecuada para implementar el algoritmo *Gradient Boosting*, se ha utilizado, sobre datos de validación, la técnica de validación cruzada. Los resultados se muestran en la Fig. 19. En ellos se aprecia que la constante de regularización $v$=0.1 es la que comete menor error cuadrático medio.

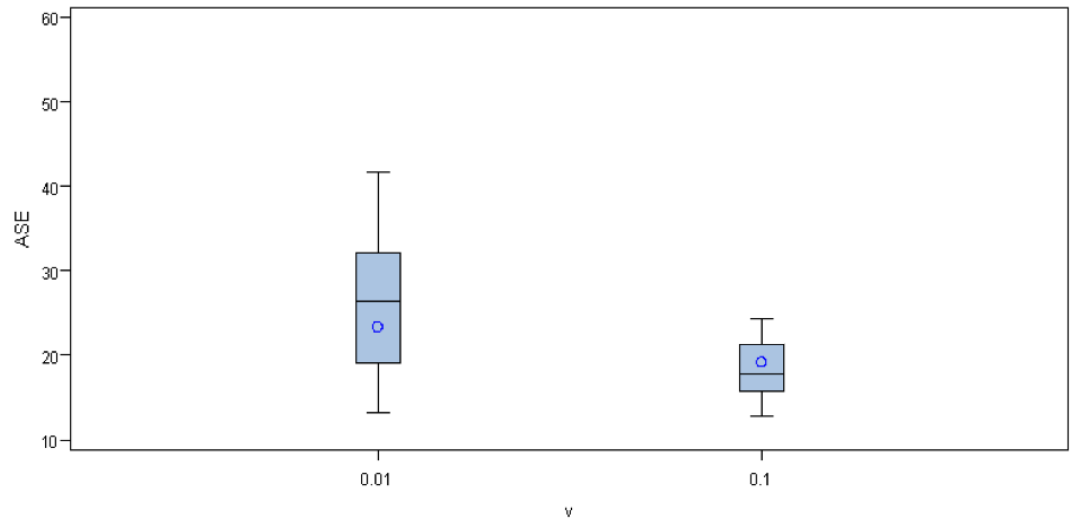


Figura 19: Gráfico donde se compara el error que se comete utilizando 40 hojas con una constante de regularizació n$v = 0{,}01$ o 40 hojas y una constante de regularización $v = 0{,}1$

Para saber si estas diferencias son estadísticamente significativas, se ha realizado un test para la media sobre muestras pareadas, cuyos resultados se muestran en el Cuadro 6.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{v=0,01}$=$\mu_{v=0,1}$ | -5.39 | $< .0001$ |

Cuadro 6: Resultados del test para la media sobre muestras pareadas utilizando constantes d egularización $v = 0{,}01$ y $v = 0{,}1$ y datos de entrenamiento.

El *p-valor* inferior a 0.001 asociado al test indica que las diferencias que se aprecian de forma gráfica son significativas, por lo que la constante de regularización adecuada para implementar el algoritmo *Gradient Boosting* es $v$=0.1.

Una vez estimados los dos parámetros requeridos para la implementación de algoritmo *Gradient Boosting*, se ha analizado la posibilidad de limitar el número de iteraciones que utiliza el algoritmo. Para ello, se ha utilizado una *macro* en SAS que permite visualizar de forma gráfica el error cuadrático medio que se comete en los datos training y en los datos de validación en función del número de iteraciones. Esta limitación conocida como *Early Stopping*, está indicada para hacer frente al sobreajuste, pues con un número muy elevado de iteraciones, podría darse el caso de que el algoritmo se ajuste a una característica de los datos.

La *macro* que se ha utilizado, además de permitir ver el comportamiento del modelo, calcula su $R^2$ y lo compara con el $R^2$ que se obtiene utilizando el modelo de regresión lineal. De esta forma, se ha tenido idea acerca del funcionamiento del algoritmo *Gradient Boosting* frente al modelo de regresión lineal.

Esta representación gráfica se puede ver en la Fig. 20, donde se han usado árboles de decisión de 40 hojas y una constante de regularización $v$=0.1.

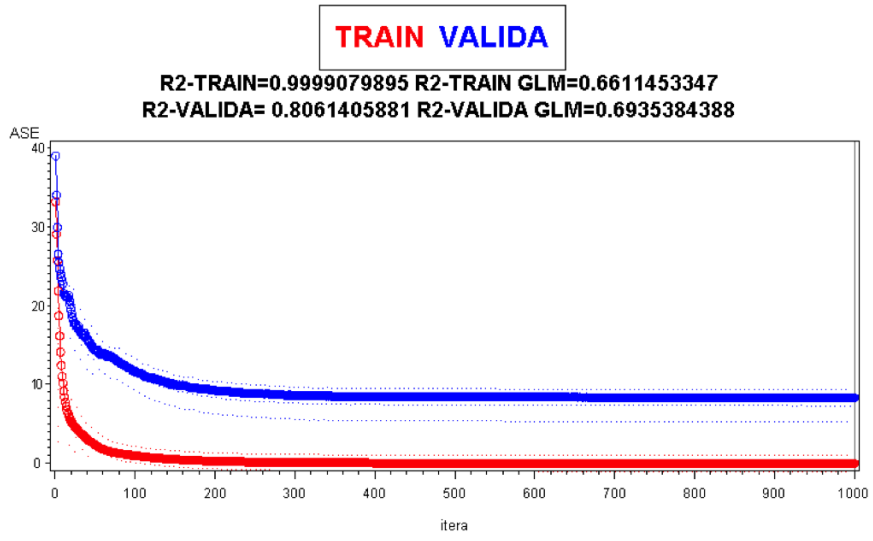


Figura 20: Comparación del error que se comete utilizando 40 hojas y una constante de regularización $v = 0{,}1$ en función del número de iteraciones que utilice el algoritmo *Gradient Boosting*.

Los resultados indican que, utilizando 100 iteraciones, el error cuadrático medio ya se encuentra estabilidazo, por lo que se ha limitado el número de iteraciones

del algoritmo a 100. En cuanto al $R^2$ que muestra la *macro*, se ve que sobre datos de entrenamiento, el algoritmo *Gradient Boosting* es capaz de explicar casi toda la variablidad de los datos. Sobre datos de validación, aunque el valor del $R^2$ baja de 0.99 a 0.80, sigue siendo superior al $R^2$ que se obtiene utilizando el modelo de regresión lineal.

## 4.5. Implementación de una *Red Neuronal*

Una vez implementado el algoritmo *Gradient Boosting* se ha implementado la *Red Neuronal*.

Para ello, SAS requiere que antes de introducir los datos se ejecute el procedimiento DMDB, que dispone los datos de forma que el módulo de minería de datos de SAS sea capaz de entenderlos.

Una vez ejecutado el procedimiento, se ha estimado el número de nodos adecuado para la implementación de la *Red Neuronal*. Para ello, se ha utilizado una *macro* en SAS que calcular el error que se comete utilizando un rango de nodos ocultos en la red, tanto para datos de entrenamiento como para datos de validación. En este caso, se ha establecido un rango de 5 a 55, obteniendo la Fig. 21.

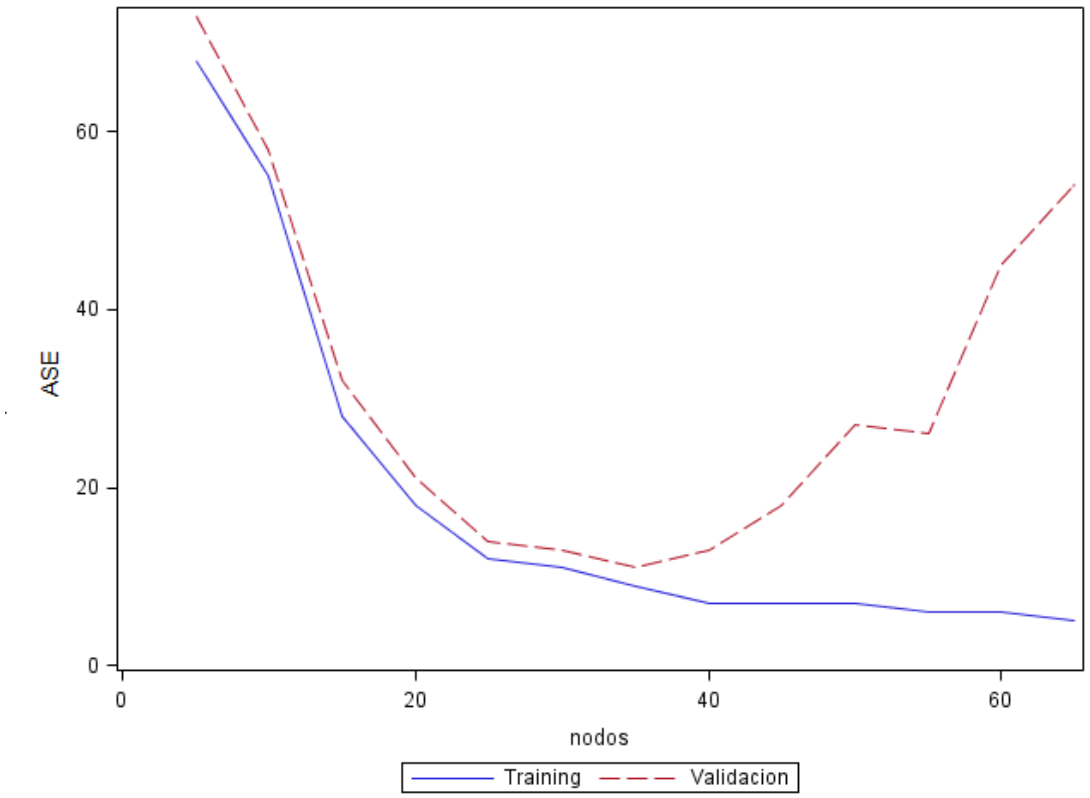


Figura 21: Comparación del error que se comete utilizando distintos nodos ocultos en la *Red Neuronal* sobre datos de entrenamiento y validación.

Se puede ver cómo el error cuadrático medio que se obtiene utilizando los datos entrenamiento o *training* disminuye de forma constante mientras que el error cuadrático medio que se obtiene utilizando datos de validación disimuye hasta que llega hasta 30 nodos, donde comienza aumentar. Tal y como se explicó en la sección 3, la curva del error cuadrático medio sobre datos de validación es la que sirve para saber el número de nodos que hay que utilizar ya que es la que muestra el punto a partir del cual el modelo comienza a estar sobreajustado.

Dado que con la Fig. 21 resulta algo difícil decidir cuántos nodos se van utilizar en la implementación de la *Red Neuronal*, existe una forma alternativa de calcularlo: la técnica de validación cruzada. En la Fig. 22 se muestran diversos diagramas de cajas llevados a cabo sobre datos *training*, uno para cada nodo, y donde se puede ver el mismo comportamiento que se veía en la línea de datos de entrenamiento de la Fig. 21.

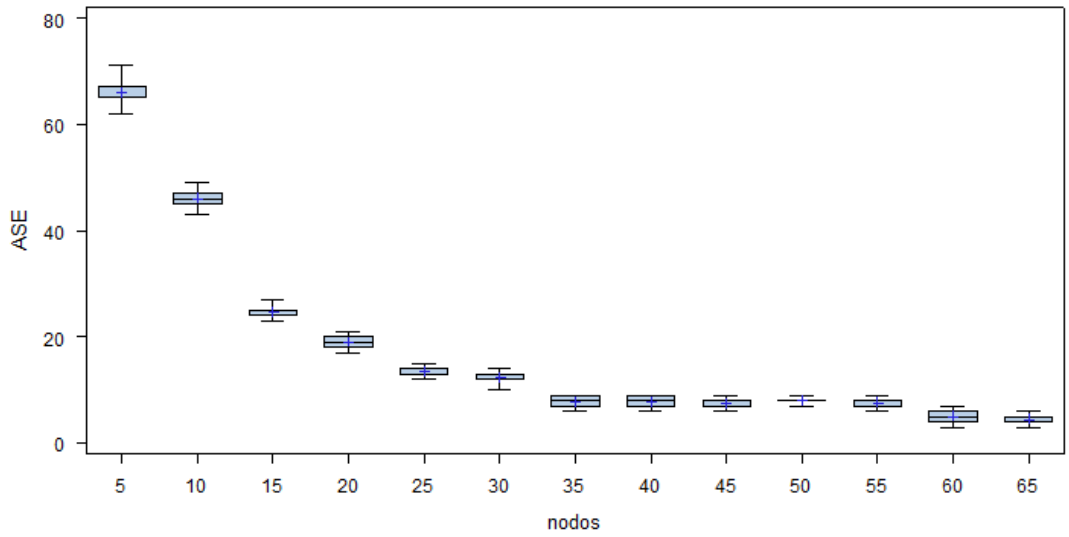


Figura 22: Error que se comete en función del número de nodos ocultos en la *Red Neuronal*, sobre datos de entrenamiento, recurriendo a la técnica de validación cruzada.

Si esta misma técnica se aplica sobre datos de validación, a través de la Fig. 23 se ve el mismo comportamiento que la línea de datos de validación en la Fig. 21, con la diferencia de que en este caso, se aprecia que el cambio de tendencia está situado en 35 nodos.

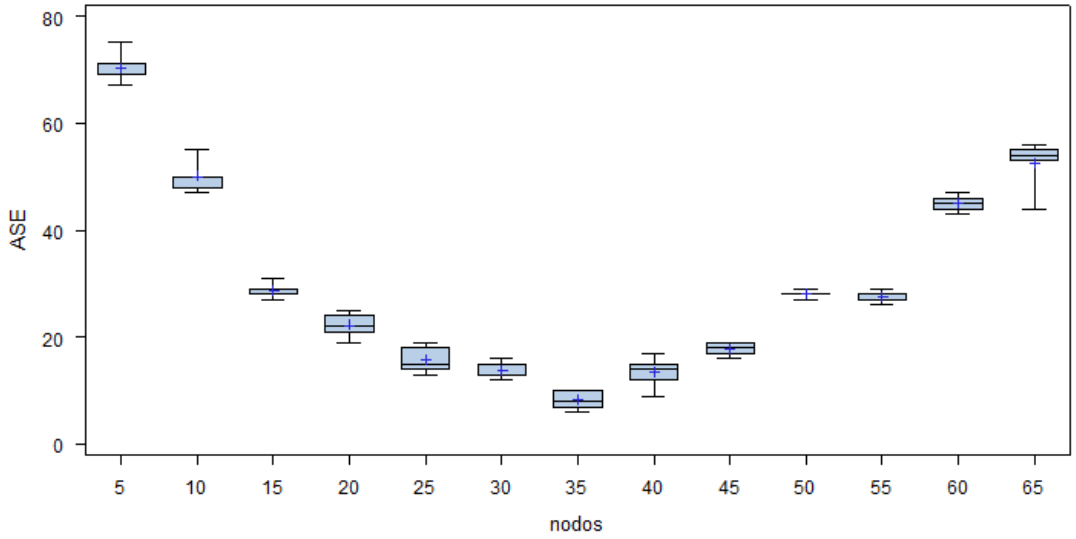


Figura 23: Gráfico donde se compara el error que se comete utilizando distintos nodos ocultos en la *Red Neuronal* utilizando la técnica de validación cruzada sobre datos de validación.

Una vez estimado el número de nodos con el que implementar la *Red Neuronal*, y al igual que se hizo en el algoritmo *Gradient Boosting* se ha analizado si es recomendable el uso o no de la limitación del número de iteraciones del algoritmo. Para ello se ha utilizado la técnica de validación cruzada, cuyos resultados se muestran en la Fig. 24. En este caso, tras llevar a cabo varias pruebas, la limitación se ha establecido 150 iteraciones, y se aprecia que la técnica de limitación de iteraciones, conocida como *Early Stopping*, disminuye el error cuadrático medio que comete la *Red Neuronal*.

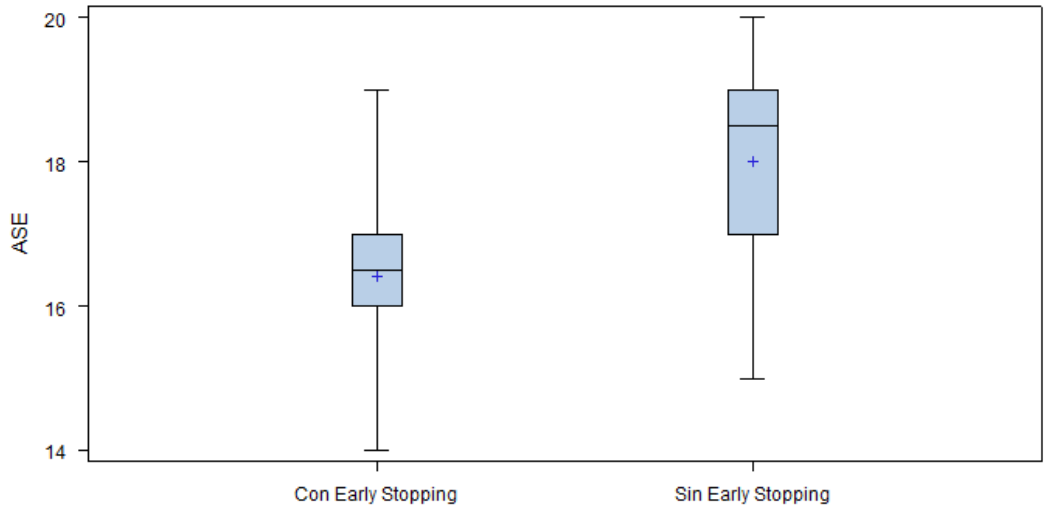


Figura 24: Error que se comete utilizando o no la limitació de 150 iteraciones.

No obstante, a pesar de que gráficamente la diferencia parece significativa, se ha llevado a cabo un test para la media sobre muestras pareadas para verificarlo.

Los resultados se muestran en el Cuadro 7 y, con un *p-valor* inferior a 0.0001, el test establece diferencias significativas a favor de la técnica de limitación de iteraciones o *Early Stopping*.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{EarlyStopping}$=$\mu_{SinEarlyStopping}$ | -11.27 | $< .0001$ |

Cuadro 7: Resultados del test para la media sobre muestras pareadas de utilizando *Early Stopping* o no

Para finalizar la implementación de la *Red Neuronal*, se ha comparado el error cuadrático medio que se comete utilizando las dos técnicas de optimización explicadas en el apartado 3: *Backpropagation* y el algoritmo de Levenberg-Mardquardt. Mediante validación se ha cuantificado el error que comete cada una de las dos técnicas. Los resultados se muestran en la Fig. 25 y se ve como la técnica *Backpropagation* es preferible al algoritmo de Levenberg-Mardquardt.

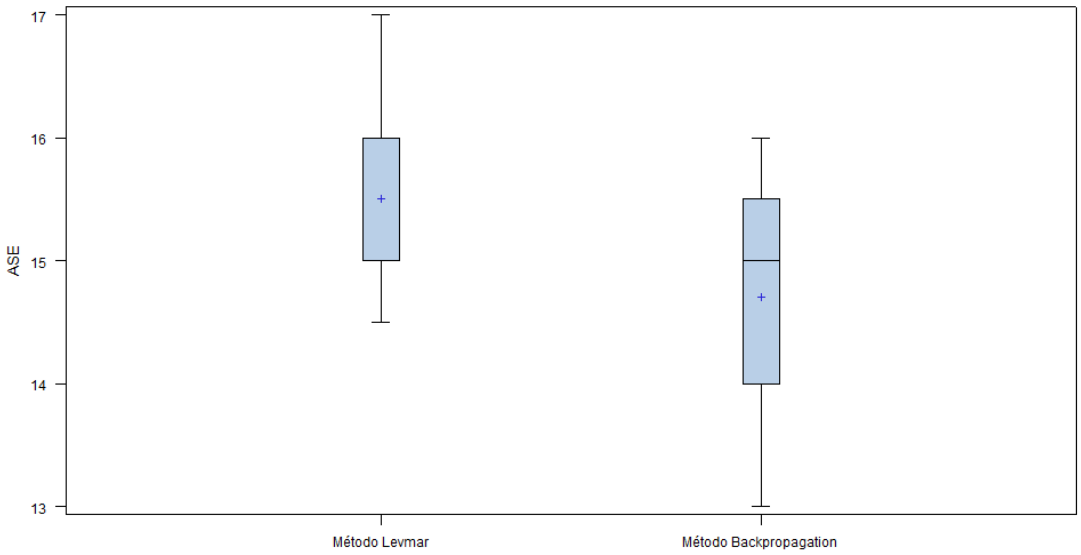


Figura 25: Gráfico donde se compara el error que se comete utilizando el algoritmo de Levenberg-Mardquardt y la técnica de optimización *Backpropagation*.

Para saber si exiten diferencias significativas entre ambas técnicas, se ha realizado un test para la media sobre muestras pareadas. Los resultados, mostrados en el Cuadro 8, establecen diferencias significativas a favor del uso de la técnica *Backpropagation*.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{Levmar}$=$\mu_{Backpropagation}$ | 8.28 | $< .0001$ |

Cuadro 8: Resultados del test para la media sobre muestras pareadas de utilizando el algoritmo de Levenberg-Mardquardt y la técnica *Backpropagation*.

## 4.6. Comparación de las técnicas de predicción

En la Fig. 26, se muestra la gráfica del error cuadrático medio que cometen las tres técnicas de predicción utilizadas: el modelo de regresión lineal, el algoritmo *Gradient Boosting* y la *Red Neuronal*. En ella, se aprecia cómo la técnica que comete el menor error cuadrático medio es la *Red Neuronal*.

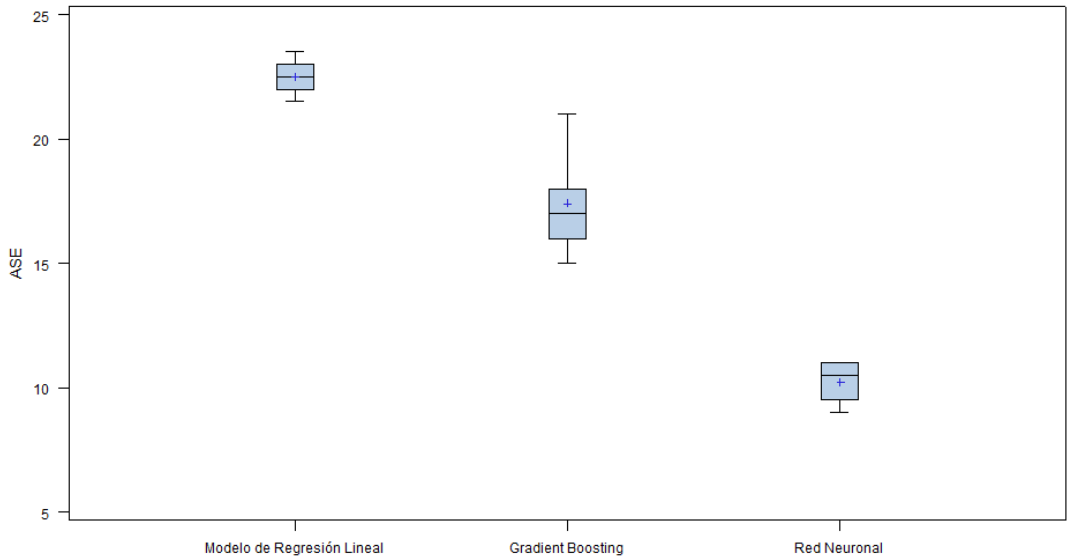


Figura 26: Gráfico donde se compara el error que se comete utilizando el modelo de regresión lineal, el algoritmo *Gradient Boosting* y la *Red Neuronal*.

Para saber si las diferencias son estadísticamente significativas, se hacemos un test para la media donde se ha comparado el modelo de regresió lineal con el algoritmo *Gradient Boosting*, el modelo de regresió lineal con la *Red Neuronal* con y el el algoritmo *Gradient Boosting* con la *Red Neuronal*. Los resultados, mostrados en el Cuadro 9, establecen diferencias significativas a favor de la *Red Neuronal*.

| $H_0$ | Estadístico $t$ | $Pr > \|t\|$ |
|---|---|---|
| $\mu_{RL}$=$\mu_{GB}$ | -14.81 | $< .0001$ |
| $\mu_{RL}$=$\mu_{RN}$ | -22.24 | $< .0001$ |
| $\mu_{GB}$=$\mu_{RN}$ | -12.37 | $< .0001$ |

Cuadro 9: Resultados del test para la media sobre muestras pareadas comparando el modelo de regresión lineal con el algoritmo *Gradient Boosting*, el modelo de regresión lineal con la *Red Neuronal* y el algoritmo *Gradient Boosting* con la *Red Neuronal*.

Estas diferencias también se pueden apreciar en las Fig. 27, 28 y 29, donde, para cada técnica de predicción utilizada, se ha calculado la mediana del error cuadrático medio cometido por hogar de cada estado, permitiendonos visualizar en cuáles se comete mayor o menor error en la predicción.

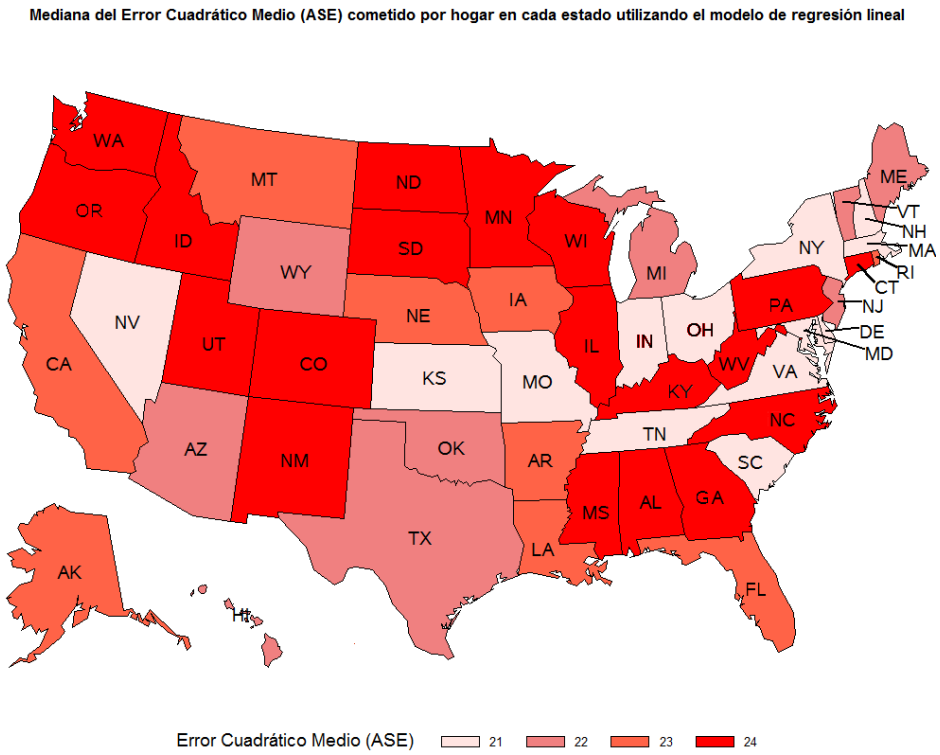


Figura 27: Mapa con la mediana del error cuadrático medio cometido por hogar en cada estado utilizando el modelo de regresión lineal.

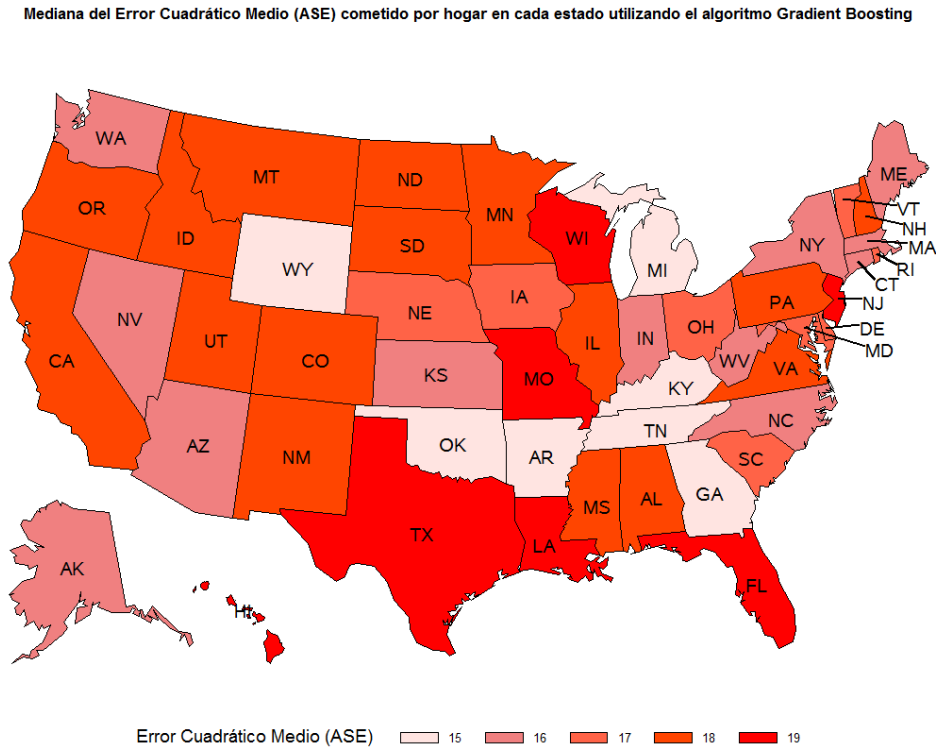


Figura 28: Mapa con la mediana del error cuadrático medio cometido por hogar en cada estado utilizando el algoritmo *Gradient Boosting*.

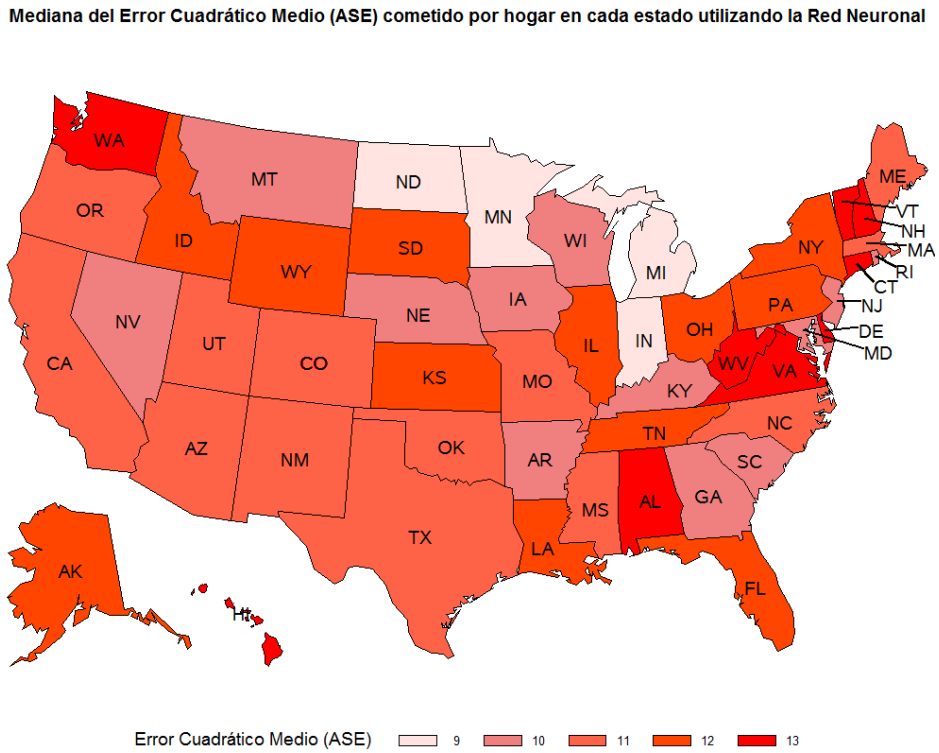


Figura 29: Mapa con la mediana del error cuadrático medio cometido por hogar en cada estado utilizando la *Red Neuronal*.

Finalmente, en las Fig. 30, 31, 32 y 33 se muestra una comparación entre la tasa de retorno real, y las tasas de retorno estimadas a través de las distintas técnicas predictivas utilizadas. La Fig. 30 contiene la mediana de la tasa de retorno real de los hogares de cada estado, mientras que las Fig. 31, 32 y 33 contienen la mediana de las estimaciones de la tasa de retorno de los hogares en cada estado. En ellas, se ve como las estimaciones obtenidas con el modelo de regresión lineal (ver Fig. 31) difieren de los valores reales provistos por el censo de los Estados Unidos (ver Fig. 30). Las estimaciónes obtenidas con el algoritmo *Gradient Boosting* (ver Fig. 32), tienen una mayor semejanza a los valores reales, aunque la mejor aproximación a se consigue utilizando la *Red Neuronal* (ver Fig. 33).

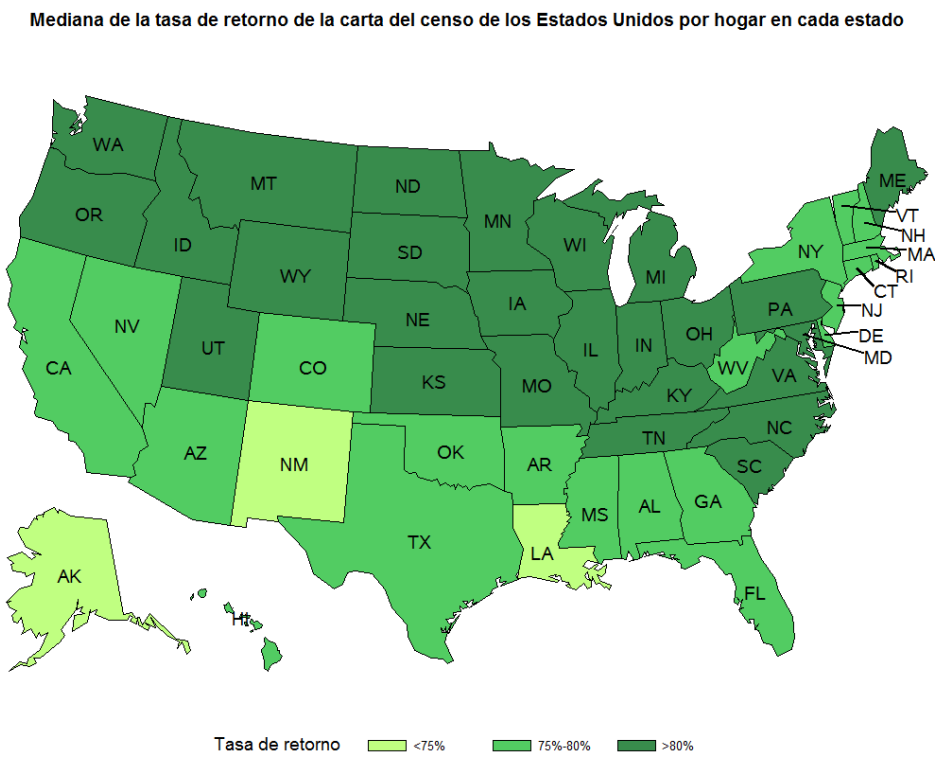


Figura 30: Mapa con la mediana de la tasa de retorno provista por el Censo de los Estados Unidos.

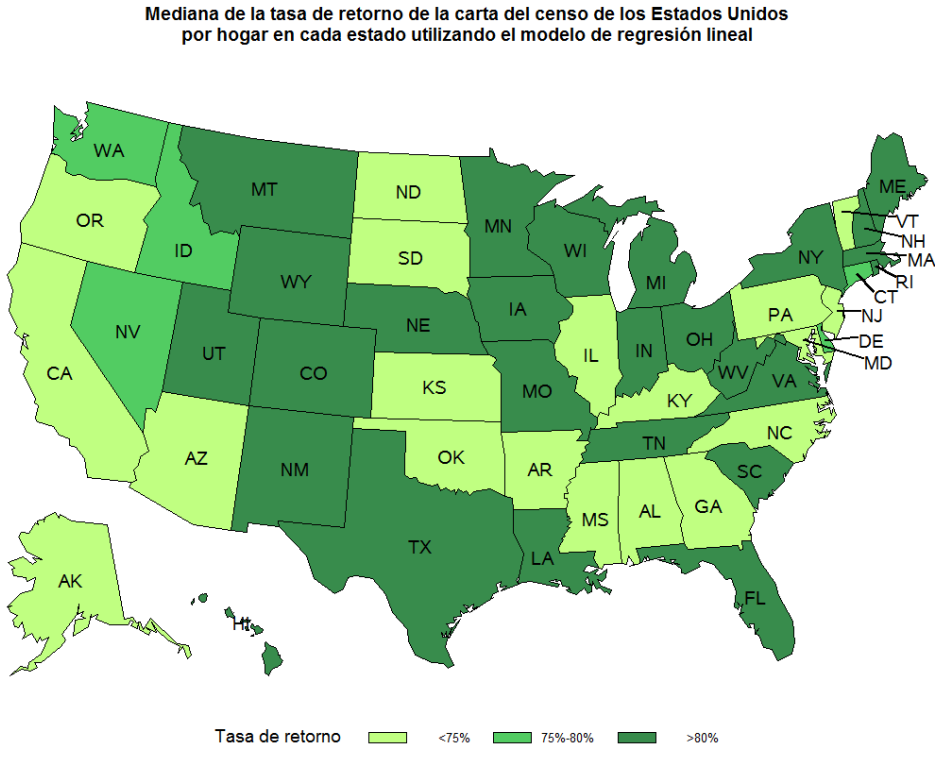


Figura 31: Mapa con la mediana de la tasa de retorno estimada de la carta del Censo de los Estados Unidos utilizando el modelo de regresión lineal.

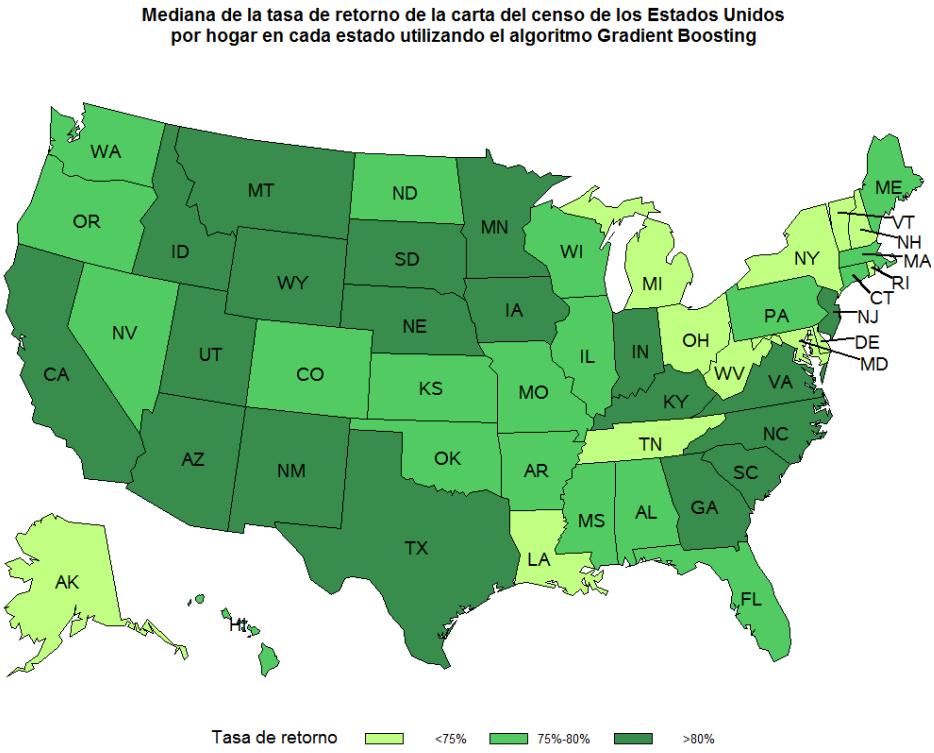


Figura 32: Mapa con la mediana de la tasa de retorno estimada de la carta del Censo de los Estados Unidos utilizando el algoritmo *Gradient Boosting*.

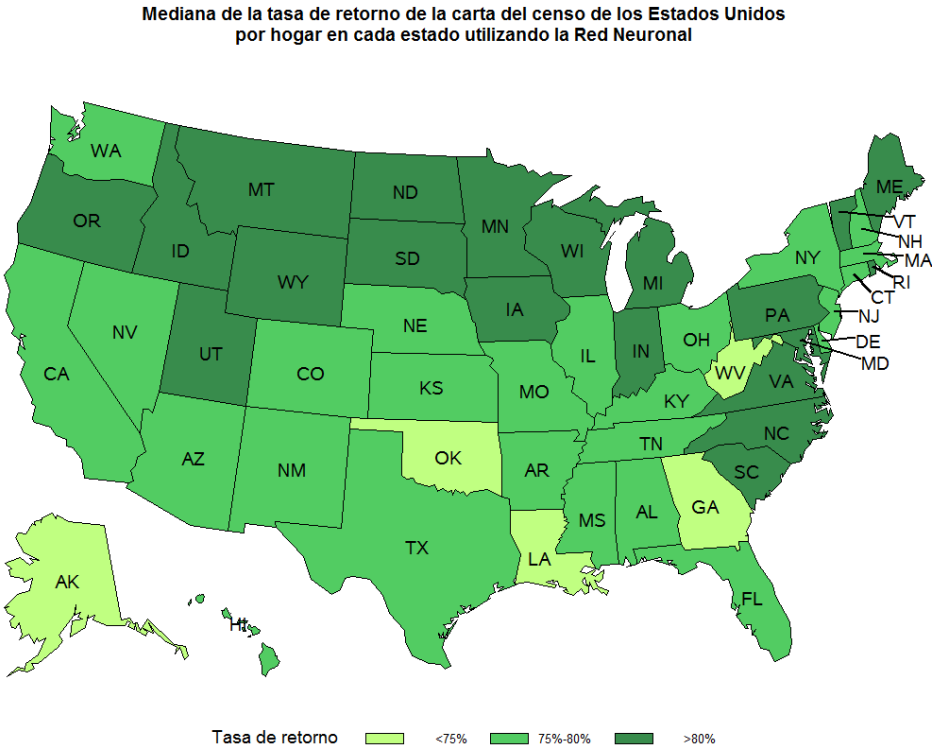


Figura 33: Mapa con la mediana de la tasa de retorno estimada de la carta del Censo de los Estados Unidos utilizando la *Red Neuronal*.

# 5. Conclusiones

En este proyecto se ha conseguido demostrar la premisa de que las técnicas de aprendizaje automático como el *Gradient Boosting* o las *Redes Neuronales*, cuando se trabaja con grandes volúmenes de información, son capaces de disminuir el error que comete un modelo de regresión lineal.

Para ello, primero se ha que depurado la base de datos de 129.605 observaciones y se han realizado las transformaciones necesarias de las 71 variables disponibles. Seguidamente, se han efectuado tres búsquedas de un modelo de regresión lineal óptimo, cada una con características distintas:

- En la primera búsqueda, las variables independietes han sido transformadas tomando el logaritmo neperiano de sus valores, y la variable dependiente ha sido transformada tomando también el logaritmo neperiano de sus valores.
- En la segunda búsqueda, todas las variables, tanto las independientes como la dependiente estaban en su estado original.
- En la tercera búsqueda, las variables independietes han sido transformadas tomando el logaritmo neperiano de sus valores, y la variable dependiente ha sido transformada tomando también la raíz cuadrada de sus valores.

De entre las 300 posibilidades que ofrecía cada búsqueda, se ha seleccionado el modelo de regresión lineal que cometiese menor error cuadrático medio de cada una de ellas. A continuación se han comparado los mejores modelos de regresión obtenidos, siendo el modelo que contenía todas las variables en su estado original el que menor error cuadático medio cometía, alcanzando así el objetivo principal del proyecto: obtener un modelo de regresión lineal que cometa el menor error posible a la hora de llevar a cabo la estimación de la tasa de retorno de la carta de la oficina del Censo de los Estados Unidos.

Una vez obtenido el modelo de regresión lineal, se ha estudiado si con el uso de otras técnicas de predicción alternativas, utilizando siempre las mismas variables, se podía reducir el error cuadrático medio que éste cometía.

En primer lugar, se ha implementado la técnica *Gradient Boosting*, que al funcionar con árboles de decisión, ha requerido estimar el número de hojas que se debían utilizar, así como la constante de regularización para evitar el sobreajuste. El resultado, utilizando 40 hojas, una constante de regularización $v$=0.1, y un límite de 100 iteraciones, ha sido una disminución en el error cuadrático medio respecto al modelo de regresión lineal.

En segundo lugar, se ha implementado una *Red Neuronal*. Una de las tareas más importantes ha sido la estimación del número de nodos que ha utilizado la red, pues al igual que en el algoritmo *Gradient Boosting*, existía riesgo de sobreajuste. Aparte del número de nodos también se ha tenido que decidir si se iba a utilizar la técnica *Early Stopping* o no, así como la técnica de optimización de la función de pérdida. El resultado, utilizando 35 nodos, la técnica de optimización

*Backpropagation* y limitanto el número de iteraciones a 150, ha sido una disminución del error cuadrático, no sólo respecto al modelo de regresión lineal, sino también respecto al algoritmo *Gradient Boosting*.

Para ver si efectivamente las técnicas de parendizaje automático eran capaces de disminuir el error que comete el modelo de regresión lineal, y saber cuál de estas técnicas es la adecuada, se ha realizado un diagrama de cajas en el que, mediante validación cruzada, se ha cuantificado el error que cometía cada una de las técnicas, concluyendo que la *Red Neuronal* es la técnica que menor error cuadrático medio comete, alcanzando así los objetivos secundarios en los que se trataba de disminuir el error que comete el modelo de regresión lineal.

Para ilustrar los resultados obtenidos, se han realizado 2 tipos de representaciones gráficas:

- El primer tipo de representaciones gráficas permite ver la mediana del error cuadrático medio cometido por familia y por estado en función de la técnica de predicción utilizada. En ellos, se puede ver la variabilidad en el error de predicción representada en el diagrama de cajas.
- En el segundo tipo de representaciones gráficas se visualiza la mediana de la tasa de retorno de la carta del censo de los Estados Unidos por hogar en cada estado. En primer lugar se ha presentado un mapa con las tasas de retorno provistas por el censo, donde se aprecia que los estados de norte poseen tasas de retorno superiores a los estados del Sur y Alaska. A continuación, se ha representado en 3 gráficos distintos, la mediana de las estimaciones de las tasas de retorno de la carta del censo de los Estados Unidos por hogar en cada estado obtenidas con cada técnica de predicción utilizada. En estos 3 gráficos se observa que las estimaciones obtenidas con el modelo de regresión lineal son las que menos se asemejan a las tasas de retorno reales, obteniendo la mayor similitud con las estimaciones obtenidas con la *Red Neuronal*.

Los resultados obtenidos en este proyecto, acompañados por las representaciones gráficas, han permitido corroborar la premisa de que las técnicas de aprendizaje automático consiguen disminuir el error de predicción que comete el modelo de regresión lineal, siendo las *Redes Neuronales* las que mejor resultado ofrecen.

# 6. Bibliografía

# Referencias


[1] Javier Portela, "Apuntes de la asignatura: Técnicas avanzadas de predicción", Universidad Complutense de Madrid, 2013.

[2] SAS Institute Inc. 2008, SAS/STAT 9.2 User's Guide, Cary, NC: SAS Institute Inc.

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

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

[5] SAS Institute Inc. 2000, SAS/Enterprise Miner The DMDB Procedure, Cary, NC: SAS Institute Inc.

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

[7] Duda, R. O. Hart, E. P. Stork, G. D. "Patter Classification". Wiley. 2001.

# Cuadernos de Trabajo
## Facultad de Estudios Estadísticos

---

**CT03/2013** **Provisión de siniestros de incapacidad temporal utilizando análisis de supervivencia.**
*Ana Crespo Palacios y Magdalena Ferrán Aranaz*

**CT02/2013** **Consumer need for touch and Multichannel Purchasing Behaviour.**
*R. Manzano, M. Ferrán y D. Gavilán*

**CT01/2013** **Un método gráfico de comparación de series históricas en el mercado bursátil.**
*Magdalena Ferrán Aranaz*

**CT03/2012** **Calculando la matriz de covarianzas con la estructura de una red Bayesiana Gaussiana**
*Miguel A. Gómez-Villegas y Rosario Susi*

**CT02/2012** **What's new and useful about chaos in economic science.**
*Andrés Fernández Díaz, Lorenzo Escot and Pilar Grau-Carles*

**CT01/2012** **A social capital index**
*Enrique González-Arangüena, Anna Khmelnitskaya, Conrado Manuel, Mónica del Pozo*

**CT04/2011** **La metodología del haz de rectas para la comparación de series temporales.**
*Magdalena Ferrán Aranaz*

**CT03/2011** **Game Theory and Centrality in Directed Social Networks**
*Mónica del Pozo, Conrado Manuel, Enrique González-Arangüena y Guillermo Owen.*

**CT02/2011** **Sondeo de intención de voto en las elecciones a Rector de la Universidad Complutense de Madrid 2011**
*L.Escot, E. Ortega Castelló y L. Fernández Franco (coords)*

**CT01/2011** **Juegos y Experimentos Didácticos de Estadística y Probabilidad**
*G. Cabrera Gómez y Mª.J. Pons Bordería*

**CT04/2010** **Medio siglo de estadísticas en el sector de la construcción residencial**
*M. Ferrán Aranaz*

**CT03/2010** **Sensitivity to hyperprior parameters in Gaussian Bayesian networks.**
*M.A. Gómez-Villegas, P. Main, H. Navarro y R. Susi*

**CT02/2010** **Las políticas de conciliación de la vida familiar y laboral desde la perspectiva del empleador. Problemas y ventajas para la empresa.**
*R. Albert, L. Escot, J.A. Fernández Cornejo y M.T. Palomo*

**CT01/2010** **Propiedades exóticas de los determinantes**
*Venancio Tomeo Perucha*

**CT05/2009** **La predisposición de las estudiantes universitarias de la Comunidad de Madrid a auto-limitarse profesionalmente en el futuro por razones de conciliación**
*R. Albert, L. Escot y J.A. Fernández Cornejo*

**CT04/2009** **A Probabilistic Position Value**
*A. Ghintran, E. González–Arangüena y C. Manuel*

**CT03/2009** **Didáctica de la Estadística y la Probabilidad en Secundaria: Experimentos motivadores**
*A. Pajares García y V. Tomeo Perucha*

**CT02/2009** **La disposición entre los hombres españoles a tomarse el permiso por nacimiento. ¿Influyen en ello las estrategias de conciliación de las empresas?**
*L. Escot, J.A. Fernández-Cornejo, C. Lafuente y C. Poza*

**CT01/2009** **Perturbing the structure in Gaussian Bayesian networks**
*R. Susi, H. Navarro, P. Main y M.A. Gómez-Villegas*

**CT09/2008** **Un experimento de campo para analizar la discriminación contra la mujer en los procesos de selección de personal**
*L. Escot, J.A. Fernández Cornejo, R. Albert y M.O. Samamed*

**CT08/2008** **Laboratorio de Programación. Manual de Mooshak para el alumno**
*D. I. de Basilio y Vildósola, M. González Cuñado y C. Pareja Flores*

**CT07/2008** **Factores de protección y riesgo de infidelidad en la banca comercial**
*J. Mª Santiago Merino*

**CT06/2008** **Multinationals and foreign direct investment: Main theoretical strands and empirical effects**
*María C. Latorre*

**CT05/2008** **On the Asymptotic Distribution of Cook's distance in Logistic Regression Models**
*Nirian Martín y and Leandro Pardo*

**CT04/2008** **La innovación tecnológica desde el marco del capital intelectual**
*Miriam Delgado Verde, José Emilio Navas López, Gregorio Martín de Castro y Pedro López Sáez*

**CT03/2008** **Análisis del comportamiento de los indecisos en procesos electorales: propuesta de investigación funcional predictivo-normativa**
*J. Mª Santiago Merino*

**CT02/2008** **Inaccurate parameters in Gaussian Bayesian networks**
*Miguel A. Gómez-Villegas, Paloma Main and Rosario Susi*

**CT01/2008** **A Value for Directed Communication Situations.**
*E. González–Arangüena, C. Manuel, D. Gómez, R. van den Brink*


UNIVERSIDAD COMPLUTENSE
MADRID