FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADÍSTICA APLICADA

Curso 2022/2023

Trabajo de Fin de Grado

***TÍTULO:* ESTUDIO PREDICTIVO DE LA DESPOBLACIÓN EN ESPAÑA.**

SUBTÍTULO: ANÁLISIS DE LA DENSIDAD POBLACIONAL EN ESPAÑA CON VARIABLES CATEGÓRICAS.

***Alumno*: ARTURO CANTERA ÁLVAREZ**

***Tutor*: PABLO ARCADIO FLORES VIDAL**

Junio (o Septiembre) de 2023

UNIVERSIDAD COMPLUTENSE
MADRID

ÍNDICE

1. **Introducción**……4
2. **Nota metodológica**……6
3. **Marco teórico**……7
   - **3.1 Asociación de variables**……7
     - 3.3.1. Coeficiente de contingencia
     - 3.3.2. V de Cramer
   - **3.2 Regresión logística ordinal**……8
     - 3.2.1. Estimación de los parámetros
   - **3.3 Evaluación del modelo**……10
     - 3.3.1. Pruebas individuales sobre los parámetros
     - 3.3.2. Matriz de confusión
     - 3.3.3. Sensibilidad y especificidad
     - 3.3.4. Pruebas de concordancia
     - 3.3.5. Precisión del modelo
     - 3.3.6. Prueba de bondad de ajuste (Pseudo R^2)
     - 3.3.7. Validación cruzada
4. **Datos empleados**……14
5. **Depuración de los datos**……15
   - **5.1 Creación de la variable dependiente**……15
   - **5.2 Depuración de variables independientes**……16
   - 5.3 **Conjunto final de datos**……22
   - **5.4 Selección de la muestra**……23
6. **Análisis descriptivo de las variables**……23
   - **6.1. Variable dependiente**……23
   - **6.2. Asociaciones entre variable dependiente e independientes**……24
   - **6.3. Variables independientes**……25
7. **Regresión logística ordinal**……28
   - **7.1. Selección de variables**……28
     - 7.1.1. Modelo completo
     - 7.1.2. Modelo final
   - **7.2. Evaluación del modelo**……32
     - 7.2.1 Pruebas individuales sobre los parámetros e interpretación.
     - 7.2.2. Matriz de confusión
     - 7.2.3. Sensibilidad y especificidad
     - 7.3.4 Pruebas de concordancia
     - 7.3.5 Precisión del modelo
     - 7.3.6. Prueba de bondad de ajuste
     - 7.3.7. Estabilidad del modelo
8. **Conclusiones**……42
   - **8.1. Posibles mejoras del estudio**……44
9. **Bibliografía**……45
10. **Anexo…**……46

1. INTRODUCCIÓN

El objetivo principal de este trabajo es el estudio de las características que condicionan la densidad de población mediante la utilización de técnicas de Machine Learning en los municipios de España a partir de las características de las viviendas y las personas residentes en ellas. Es decir, se tratará de modelar el problema de la despoblación, con el objetivo de encontrar factores que condicionan el crecimiento demográfico de zonas poco habitadas.

A raíz de ello, definimos el término de la despoblación, siendo la pérdida total o parcial de los habitantes de un lugar. Se considera "zonas con muy baja densidad de población: regiones NUTS 3 con menos de 12,5 habitantes por km2." (COMISIÓN EUROPEA, 2013). En España, NUTS3 corresponde con cada una de las provincias, Ceuta y Melilla.

Actualmente la despoblación es un gran problema en España, pues "hay 2.936 municipios con una densidad inferior a los 12,5 hab/km^2 … y se extienden por 243.000 km^2, el 48% de la superficie del país" (Vicepresidencia Cuarta y Ministerio para la Transición Ecológica y el Reto Demográfico, 2020), como podemos ver en el siguiente mapa, donde los municipios con menos de 12,5 habitantes por km^2 en 2019 aparecen en color verde.

Ilustración 1: Municipios con densidad de población inferior a 12,5 habitantes/km2 en 2019

Este problema es uno de los principales desafíos a los que se está enfrentando la Unión Europea, pues "observa que las zonas rurales y las regiones intermedias representan el 88 % del territorio de la UE; en ellas vive el 55 % de la población, se genera el 43 % de su valor añadido bruto y está implantado el 56 % de sus puestos de trabajo. También observa que las zonas rurales no son homogéneas, sino que algunas de ellas están afectadas por importantes desafíos demográficos (despoblación, problemas de envejecimiento, etc.) que impiden su desarrollo económico y social. Por lo tanto, el desarrollo rural es de extrema importancia para el Comité de las Regiones (CDR) y un instrumento vital para alcanzar el objetivo de cohesión territorial consagrado en el Tratado de Lisboa" (Presidente del Comité Europeo de las Regiones, 10 de Diciembre de 2020).

Por otro lado, se define Machine Learning como un método de IA que se utiliza cada vez más para el procesamiento de grandes datos. Básicamente es que un programa de computadora puede aprender y adaptarse a nuevos datos sin participación humana. Asimismo, para que el aprendizaje sea bueno, preciso y efectivo, se necesitan de grandes volúmenes de datos. (Mehryar Mohri, 2018).

2. NOTA METODOLÓGICA

El proceso por seguir para la realización del estudio será la siguiente:

En primera instancia se explicará el marco teórico de las diferentes técnicas estadísticas a emplear, así como todos los conceptos necesarios para la correcta comprensión del propio estudio.

A continuación, se expondrán los datos disponibles para el estudio y el proceso llevado a cabo para el tratamiento y depuración previo. Específicamente, se tratará la creación de la variable dependiente, la depuración del resto de variables y el método de selección de la muestra con la que se llevará a cabo el posterior análisis.

En tercer lugar, se llevará a cabo un análisis descriptivo de las variables más relevantes en el estudio, comenzando por la variable dependiente y continuando por las independientes más importantes, ya sea por su alta correlación con la variable dependiente o por la importancia desde un punto de vista más subjetivo para el modelado. Todo ello se realizará con el objetivo de conocer cómo se distribuyen las variables de las bases de datos anteriormente mencionadas, analizando también la correlación de las variables independientes con la dependiente.

Posteriormente se creará, aplicando la técnica de Machine Learning regresión logística ordinal, un modelo con el que predecir la clasificación de las observaciones en las categorías de la variable dependiente llevando a cabo una selección previa de variables. Se tratará de encontrar el modelo más sencillo sin pérdida de información, con el objetivo de conocer la situación poblacional del municipio en el que se encuentra una vivienda a partir de las características que presenta, sin tener ningún tipo de información geográfica de éste. También se estudiarán las métricas del modelo para determinar si el modelo es válido y de calidad desde un punto de vista estadístico.

Por último, se expondrán las conclusiones obtenidas tras el estudio y se propondrán posibles líneas para la continuación, mejora y profundización del análisis.

Todo el análisis, tanto descriptivo como regresivo será realizado con R como lenguaje de programación y como editor y herramienta, RStudio.

Este estudio está limitado al desconocer el código de los municipios con menos de 50.000 habitantes del país, dado que el INE no los ofrece en la Encuesta utilizada por secreto estadístico, problema que se tratará a continuación.

3. MARCO TEÓRICO

3.1. ASOCIACIÓN DE VARIABLES

El análisis de asociación de variables consiste en utilizar técnicas estadísticas para estudiar la existencia de algún tipo de relación entre dos o más variables para detectar la presencia de patrones o tendencias de emparejamiento entre los valores de estas.

3.1.1 COEFICIENTE DE CONTINGENCIA

"El coeficiente de contingencia de Pearson expresa la intensidad de la relación entre dos (o más) variables cualitativas. Se basa en la comparación de las frecuencias calculadas de dos características con las frecuencias que se hubiesen esperado con independencia de estas características" (Bortz J., Lienert G. A. & Boehnke K. 2016; Zofel P. 2017).

Su fórmula viene dada por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \qquad 0 \leq C \leq 1$$

Con (número de filas – 1), multiplicado por (número de columnas – 1) grados de libertad.

Cuanto más próximo de 1 se encuentre, mayor relación existe entre las variables analizadas, y por contrapartida cuanto más cercano a 0, menor relación.

Este coeficiente presenta un inconveniente y es que no muestra la dirección en la relación de las variables (solamente toma valores positivos entre 0 y 1), solamente indica la intensidad de ésta.

3.1.2 V DE CRAMER

"Funciona como una medida de relación estadística basada en Ji cuadrado. Es decir, para hacer una corrección del coeficiente Ji Cuadrado donde se pueda precisar la fuerza de asociación entre dos o más variables. En este sentido, el resultado del coeficiente varía entre cero y uno (siendo cero un valor nulo de asociación)." (Isea, Ojeda, Fernandez, Gutierrez, & Salazar, 2018).

Al igual que en el coeficiente de contingencia, su uso está limitado a variables categóricas y siempre positivo, por lo que no se puede conocer por medio de esta técnica la dirección de la relación existente entre las variables estudiadas.

La fórmula para el cálculo viene dada por:

$$V = \sqrt{\frac{\chi^2}{n(\min[r,c]-1)}}$$

Con n número total de frecuencia, y min el número menor de categorías entre filas (r) y columnas (c) menos 1.

Un valor por encima de 0.3 implica una correlación significativa.

3.2. REGRESIÓN LOGÍSTICA ORDINAL.

La regresión logística es un algoritmo de Machine Learning de clasificación en el cual se emplea para la predicción de las probabilidades de pertenencia a las categorías de una variable categórica. El resultado del análisis será un modelo logístico en el que se estiman las probabilidades de respuesta a cada una de las categorías de la variable explicada.

El método de regresión logística ordinal se emplea cuando la variable de interés tiene varias categorías. Podría crearse el modelo a partir de regresión logística multinomial, pero no se tendría en cuenta la componente ordinal de los datos, lo que supondría una pérdida de información.

La metodología de regresión logística ordinal inicia suponiendo una variable cualitativa Y con sus categorías ordenadas $y_1, y_2, \dots, y_k$. Un objetivo de la modelación es tratar de explicar el comportamiento de la variable Y mediante las variables independientes $X_1, X_2, \dots, X_m$.

"Esta variable Y se construye a partir de una ecuación de regresión lineal mediante las m características de los decisores, convertidos en variables *dummies*:" (Calviño, 2022)

$$\beta_0 + \beta_1 + \cdots + \beta_k + \varepsilon$$

Lo que equivale a que la variabe dependiente venga dada por:

$$Y = j, si\ \alpha_{j-1} < \beta_1 X_1 + \cdots + \beta_m X_m + \ \varepsilon \leq \alpha_j, j = 1, \dots, k, con\ \alpha_0 = -\infty\ y\ \alpha_k = \infty$$

Por tanto, la probabilidad de cada alternativa se puede obtener como

$$p_{ij} = P(Y = j | x_{1i}, \dots, x_{mi}) = P(\alpha_{j-1} < \beta_1 X_1 + \cdots + \beta_m X_{mi} + \ \varepsilon_i \leq \alpha_j)$$

$$= \ P(\alpha_{j-1} - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi}) < \varepsilon_i \leq \alpha_j - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi}))$$

$$= F_\varepsilon\ (\alpha_j - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi})) - = F_\varepsilon\ (\alpha_{j-1} - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi}))$$

$$= \frac{e^{\alpha_j - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi})}}{1 + e^{\alpha j - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi})}} - \frac{e^{\alpha_{j-1} - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi})}}{1 + e^{\alpha_{j-1} - (\beta_1 X_{1i} + \cdots + \beta_m X_{mi})}}$$

Donde:

- Y es el valor de la variable dependiente que se quiere predecir
- $X_1, \dots, X_m$ son los valores de las variables independientes.

- β_0 es el valor de la constante del modelo.
- $\beta_1 \dots \beta_k$ corresponde con el valor de los coeficientes de regresión a estimar para cada una de las variables regresoras.
- ε sería el término de error del modelo.

En R, se calculará con la función *polr* de la librería *MASS*.

3.2.1 ESTIMACIÓN DE LOS PARÁMETROS

Se basa en el método de máxima verosimilitud, asumiendo distribución multinomial en la variable dependiente, cuyo primer parámetro es igual a 1.

Por ello, los parámetros β_i, i = 1, . . . , m y α_i, i = 1, . . . , K – 1 serán aquellos que maximicen la siguiente función de verosimilitud:

$$L(\beta) = \prod_{i=1}^{n} p_{1i}^{y1i} p_{2i}^{y2i} \dots p_{ki}^{yki}$$

Con $p_{ij} = P(Y = j|x_{1i}, \dots, x_{mi}), \forall i = 1, \dots, n; j = 1, \dots, K$ y

$$y_{ij} = \begin{cases} 1 & si\ \ y_i = j \\ 0 & en\ otro\ caso \end{cases}, \forall i = 1, \dots, n; j = 1, \dots, K$$

3.2.3 INTERPRETACIÓN DE LOS PARÁMETROS

Para la interpretación de los parámetros de un modelo de regresión logística ordinal se recurre al concepto de los odds ratio, dado por:

$$odds(Y > j|x_{1i}, \dots, x_{mi}) = \frac{P(Y > j|x_{1i}, \dots, x_{mi})}{(Y \leq j|x_{1i}, \dots, x_{mi})} = \frac{\frac{1}{1 + e^{\alpha_j - \beta_1 x_{1i} - \dots - \beta_m x_{mi}}}}{\frac{e^{\alpha_j - \beta_1 x_{1i} - \dots - \beta_m x_{mi}}}{1 + e^{\alpha_j - \beta_1 x_{1i} - \dots - \beta_m x_{mi}}}}$$

Por tanto, el odds ratio asociado al aumento unitario de la primera variable se deduciría como:

$$\frac{odds(Y > j|(1 + x_i), \dots, x_{mi}}{odds(Y > j|x_{1i}, \dots, x_{mi}} = \frac{e^{-\alpha_j + \beta_1(1 + x_{1i}) + \dots + \beta_m x_{mi}}}{e^{-\alpha_j + \beta_1 x_{1i} + \dots + \beta_m x_{mi}}} = e^{\beta_1}$$

3.3 EVALUACIÓN DEL MODELO

Para determinar un correcto ajuste del modelo a los datos, se ha de comprobar que el modelo es adecuado, lo cual se puede comprobar a través de varias métricas y diferentes puntos.

3.3.1 PRUEBAS INDIVIDUALES SOBRE LOS PREDICTORES

El primer paso para la evaluación del modelo de regresión logística ordinal es la prueba de significancia de los estimadores de los coeficientes de los regresores. La prueba que nos da esta información es el Test de Wald.

El Test de Wald contrasta la hipótesis nula de que el coeficiente de regresión asociado a una variable predictora es igual a cero (no tiene un efecto significativo en las categorías ordenadas de la variable respuesta), mientras que la hipótesis alternativa defiende que no es igual a cero y sí tiene un efecto significativo en la variable respuesta.

$$W = \left(\frac{Coeficiente\ de\ Regresión}{Error\ Estándar\ del\ Coeficiente\ de\ Regresión}\right)^2$$

Se compara con una χ^2 de 1 grado de libertad.

Si se acepta la hipótesis nula, significa que el parámetro es significativo en el modelo.

En el software R lo visualizaremos mediante la función *stargazer* de la librería del mismo nombre *stargazer*, que permite crear tablas resumen de los modelos estadísticos (en este caso regresión logística).

3.3.2 MATRIZ DE CONFUSIÓN

La matriz de confusión en Machine Learning es una matriz que permite representar el número de predicciones de cada clase en un modelo estadístico con una variable respuesta categórica. Permite observar a simple vista los aciertos y fallos en la clasificación, así como calcular la precisión, índice Kappa. Sensibilidad y especificidad del modelo.

En el software la obtendremos a través de *confusionMatrix* de la librería *caret*.

3.3.3 SENSIBILIDAD Y ESPECIFICIDAD

Estas dos métricas se utilizan en problemas de clasificación y miden la capacidad del modelo para identificar los verdaderos positivos o negativos predichos por el modelo. (Susi, 2022)

- **Sensibilidad**

La sensibilidad, a menudo denominada "tasa de verdaderos positivos" o "tasa de detección" se refiere a la capacidad del modelo para identificar correctamente los casos positivos de la clase

objetivo, es decir, su habilidad para detectar verdaderos positivos (VP) en relación con el total de casos positivos verdaderos presentes en la población.
La fórmula para calcular la sensibilidad es la siguiente:

$$Sn = \frac{VP}{VP + FN}$$

Donde:

- VP: Número de verdaderos positivos (casos correctamente clasificados como positivos).
- FN: Número de falsos negativos (casos incorrectamente clasificados como negativos).

La sensibilidad se expresa como un valor entre 0 y 1, donde un valor más cercano a 1 indica una alta capacidad del modelo para detectar correctamente casos positivos. En otras palabras, una sensibilidad alta implica que el modelo minimiza la probabilidad de perder casos positivos importantes.

Especificidad

La especificidad, conocida como "tasa de verdaderos negativos" (Sp). Se refiere a la capacidad del modelo para identificar correctamente los casos negativos de una clase objetivo, es decir, su habilidad para detectar verdaderos negativos (VN) en relación con el total de casos negativos verdaderos presentes en la población.
La fórmula para calcular la especificidad es la siguiente:

$$Sp = \frac{VN}{VN+FP}$$

Donde:

- VN: Número de verdaderos negativos (casos correctamente clasificados como negativos).
- FP: Número de falsos positivos (casos incorrectamente clasificados como positivos).

Al igual que la sensibilidad, la especificidad se expresa como un valor entre 0 y 1, donde un valor más cercano a 1 indica una alta capacidad del modelo para identificar correctamente casos negativos. Una alta especificidad implica que el modelo minimiza la probabilidad de clasificar incorrectamente casos negativos como positivos.

3.3.4 PRUEBA DE CONCORDANCIA

Debido a la naturaleza ordinal de los datos, el índice más adecuado es el índice de Kappa ponderado, que evalúa los aciertos del modelo eliminando aquellos que se pueden producir al azar teniendo en cuenta todas las celdas de la matriz de confusión ponderadas en función de su proximidad a la diagonal principal. Para calcularlo, se hará uso de la función *Kappa* de la librería *vcd* de R.

Como en el índice Kappa clásico, partimos de la matriz de confusión relativa en la que las celdas p_{ij}representan la proporción de observaciones que pertenecen a la categoría i pero son clasificados en la j, y $p_{i.}\, p_{.j}$ son las proporciones de observaciones en la categoría i real y predicha, respectivamente. Así mismo, se definen los pesos w_{ij} para cada una de las celdas de la matriz de confusión, que tomarán valores entre 0 y 1, donde el 1 representa el mayor acierto y el 0, el mayor error. El índice Kappa ponderado viene dado por:

$$k_p = \frac{\Sigma_{i=1}^{k}\Sigma_{j=1}^{k} w_{ij} p_{ij} - \Sigma_{i=1}^{k}\Sigma_{j=1}^{k} w_{ij} p_{i.} p_{.j}}{1 - \Sigma_{i=1}^{k}\Sigma_{j=1}^{k} w_{ij} p_{i.} p_{.j}}$$

Siendo $w_{ij} = 1 - (|i - j|)/(K - 1)$

En cuanto a la calidad, se puede resumir en la siguiente tabla:

ÍNDICE KAPPA	INTERPRETACIÓN
0	Equivalente al azar
0.01-0.20	Pobre
0.21-0.40	Justo
0.41-0.60	Moderado
0.61-0.80	Bueno
0.81-1	Excelente

Tabla 1: Interpretación del Índice de Kappa

Y las ponderaciones serían las siguientes para este caso por tener 4 categorías en la variable respuesta:

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000000	0.6666667	0.3333333	0.0000000
[2,]	0.6666667	1.0000000	0.6666667	0.3333333
[3,]	0.3333333	0.6666667	1.0000000	0.6666667
[4,]	0.0000000	0.3333333	0.6666667	1.0000000

TABLA 2: Ponderaciones de Kappa ajustado

3.3.5 PRECISIÓN DEL MODELO

Se evaluará la precisión del modelo a través de la función *confusionMatrix* de la librería *caret*, donde se evaluará la precisión (accuracy) del modelo.

Esta métrica evalúa el rendimiento del modelo de clasificación. Mide la proporción de predicciones correctas realizadas por el modelo en relación con el número total de predicciones, basándose en la fórmula:

$$Precisión = \frac{Total\ de\ predicciones\ correctas}{Total\ de\ predicciones}$$

Donde 1 sería una precisión perfecta y 0, precisión nula.

3.3.6 PSEUDO-R^2

La idea detrás del pseudo-R^2 es proporcionar una medida de cuánto mejor se ajusta un modelo en comparación con un modelo nulo o base. En otras palabras, mide la mejora en el ajuste del modelo en relación con un modelo que solo incluiría una constante o una línea recta horizontal en el caso de regresión logística.

Por tanto, para este análisis dado el carácter ordinal de los datos se hará uso de la R^2 de McFadden, cuya fórmula es:

$$R^2_{McFadden} = 1 - \frac{\log\left(L_{modelo\ ajustado}\right)}{\log\left(L_{modelo\ nulo}\right)}$$

La R^2 de McFadden es la reducción proporcional en valor absoluto de log-likelihood y mide cuánto del error del ajuste disminuye al incluir las variables predictoras. Proporciona una medición de la significación real del modelo. Esta puede variar entre 0 (indicando que los predictores son inútiles prediciendo la variable respuesta) y 1 (indicando que el modelo predice perfectamente la respuesta).

No es una medida de la varianza explicada como el R^2 en modelos de regresión lineal, sino que es una medida de la bondad de ajuste específica para el tipo de modelo logístico ordinal en el que nos encontramos. Para ello se utilizará la librería *DescTools* y la función *PseudoR2* de R, con la opción *which="McFadden"*.

3.3.7 VALIDACIÓN CRUZADA

La validación cruzada es una técnica utilizada en Machine Learning para el análisis de la estabilidad de un modelo que mediante particiones, en este caso de dos, mediante la división los datos en entrenamiento-prueba para comprobar si las pruebas de precisión y concordancia analizadas anteriormente son similares en ambos conjuntos de datos, con el objetivo de descartar un sobreajuste del modelo a los datos de entrenamiento.

Por tanto, el proceso consistiría en primer lugar de la división de los datos en entrenamiento (80% de los datos) y prueba (20% restante), después se crea el modelo anteriormente expuesto con los datos de entrenamiento, se crea el mismo modelo con los datos de prueba y se comparan los resultados de las pruebas de ambos modelos. Si las medidas son similares, estaremos ante un modelo estable que no tiende al sobreajuste con los datos de entrenamiento.

4. DATOS EMPLEADOS.

Para llevar a cabo el estudio se han utilizado tres ficheros de datos, uno del Instituto Geográfico Nacional y los otros dos del Instituto Nacional de Estadística.

En primer lugar, se utiliza la Encuesta de Características Esenciales de la Población y las Viviendas (INE, 2021), una encuesta de tipo estructural de periodicidad quinquenal con información recogida desde el 29 de marzo de 2021 al 15 de febrero de 2022 a nivel nacional mediante un muestreo aleatorio estratificado y muestreo bietápico estratificado, de la cual se analizará la información de dos conjuntos de datos simultáneamente contenidos en esta encuesta: las características de la vivienda, que posee datos tales como información física de la propia vivienda (tamaño, tipo de vivienda, etc.), infraestructuras de las que dispone (fibra óptica, gas por tubería, tipo de calefacción, etc.), servicios (tener próximo a la vivienda un colegio, centros de salud, farmacias, etc.), accesibilidad y la información correspondiente al municipio donde se sitúa el domicilio, (exceptuando las localidades con menos de 50.000 habitantes, de las que solo se conoce la provincia en la que se encuentran); y por otro lado la información propia a la situación socioeconómica de las personas mayores de 18 años residentes en cada una de las viviendas encuestadas.

Posteriormente se profundizará más en las variables escogidas para el análisis dado el alto número de variables existente entre ambas bases de datos. Cabe destacar que no se utilizan todas las variables presentes en estos dos conjuntos de datos, pues algunas variables no son de interés en este estudio.

Además, se utiliza otro conjunto de datos con la densidad de población por municipios del Instituto Geográfico Nacional del año 2021, cuya información de interés es el código de la provincia (CPRO), del municipio (CODMUN-INE) y la densidad de población del mismo (DENS_POB)

Por ello, se dispone originalmente de 61 variables y 361.935 observaciones en total de las 50 provincias de España y las dos ciudades autónomas de Ceuta y Melilla. Al ser tan alta la cantidad de observaciones, se realizará previamente un muestreo aleatorio simple, condicionado por el tamaño del municipio para reducir la cantidad y que sea posible el análisis de datos desde un punto de vista computacional cuyo proceso se explicará más adelante.

5. DEPURACIÓN DE LOS DATOS

5.1 CREACIÓN DE LA VARIABLE DEPENDIENTE.

Para identificar el estado del nivel de población en el que se encuentra cada municipio, se toma como referencia el límite marcado por la Comisión Europea y la media de la población de los municipios de España para crear una nueva variable, llamada DESPOBLACIÓN que contará con 4 niveles. Esta información se obtiene, como se ha mencionado anteriormente, de los datos ofrecidos por el Instituto Geográfico Nacional para cada municipio de cuya base de datos se utilizará el código del municipio (será la variable identificadora para cruzar posteriormente los datos) y la densidad de población de cada uno.

Dado que en la Encuesta de Características Esenciales de la Vivienda y las Personas por secreto estadístico no se ofrece el código de municipio para los que tienen menos de 50.000 habitantes, se ha realizado previamente una estimación de las densidades de los municipios en cada provincia, calculando la media de las densidades de población de las localidades de cada provincia con menos de 50.000 habitantes. Es decir, para cada municipio de más de este número de habitantes se utiliza su propia densidad y para los que tienen menos se utiliza la media de la densidad de población de su provincia sin tener en cuenta la de las localidades de las que sí que es posible conocer su información poblacional.

Los niveles de la nueva variable serían los siguientes:

-X1: Municipio en despoblación (con menos de 12,5 habitantes por km^2)
-X2: Municipio en riesgo de despoblación (densidad de población entre 12,5 y 94 habitantes por km^2)

-X3: Municipio sin riesgo de despoblación (densidad entre la media de España y el doble de la media)
-X4: Municipio con alta densidad de población (densidades por encima del doble)

5.2 DEPURACIÓN DEL RESTO DE VARIABLES

Las variables independientes son propias de la Encuesta de Características Esenciales de la Población y las Viviendas. Originalmente se compone de 80 variables por parte de las características de la vivienda y 43 de las características de las personas, siendo un total de 123 variables con una observación por cada vivienda registrada sumando un total de 172.445 observaciones por vivienda y posteriormente un registro por persona de cada vivienda, haciendo un total de 361.934 observaciones.

Para este análisis se han eliminado previamente algunas variables sin interés para este estudio, como, por ejemplo, el tipo de bombillas que se utilizan en el domicilio.

Se han adaptado en función de las características de cada variable, siendo estas categóricas (variables factor) y ordinales. Además, en la variable del número de vehículos (NVEHICULOS) se ha sustituido el valor faltante por el valor 0 que correspondería con ninguno, al igual que en el tipo de calefacción (TIPOCOMBCALE) no disponer de ningún tipo de calefacción y el número de plazas de garaje (NPLAZASGAR) al igual. También se ha transformado la variable NHIJOS, convirtiendo NA en 0 que correspondería a la no tenencia de hijos.

La codificación de las variables (ordenadas alfabéticamente para cada tipo) que se utilizan para el estudio tras la depuración es:

- Codificadas como

1	Sí
6	No

Tabla 3: Codificación 1 Sí 6 No

ADAPTADA (la vivienda está adaptada a las necesidades del envejecimiento de las personas)
AGUACALCENT (El edificio tiene agua caliente central)
AIREACOND (Dispone de algún sistema de refrigeración (aire acondicionado, aparatos móviles …; NO ventiladores)
AISLAM (La vivienda tiene algún problema de aislamiento)

BARES (La zona en la que está ubicada la vivienda tiene servicios de restauración)
COLEGIO (La zona en la que está ubicada la vivienda tiene colegios)
CONTAMIN (La zona en la que está ubicada la vivienda tiene contaminación o malos olores)
CSALUD (La zona en la que está ubicada la vivienda tiene hospitales, centros de salud y/o ambulatorios)
DELINCUENCIA (La zona en la que está ubicada la vivienda tiene delincuencia o vandalismo)
ENERENOV (El edificio tiene instalación de dispositivo de energía renovable)
EVACUAGUARES (El edificio tiene sistema de evacuación de aguas residuales)
FARMACIA (La zona en la que está ubicada la vivienda tiene farmacias)
FLEXI (Puede flexibilizar, adaptar o acomodar su jornada laboral por conciliación: 1 SI, 6
GASTUBERIA (El edificio dispone de gas por tubería)
INTERNET (Dispone la vivienda de acceso a internet)
MALCOMUNIC (La zona en la que está ubicada la vivienda tiene malas comunicaciones)
MOLESTURIST (La zona en la que está ubicada la vivienda tiene molestias relacionadas con actividades turísticas y/o locales de hostelería)
PAREJA (Tiene pareja actualmente)
POCOVERDE (La zona en la que está ubicada la vivienda tiene pocas zonas verdes)
RUIDOS (La zona en la que está ubicada la vivienda tiene ruidos exteriores)
SEGUNRESI (Dispone de una segunda residencia en propiedad)
NO)
SMART (Dispone de teléfono móvil smartphone)
SUCIO (La zona en la que está ubicada la vivienda tiene poca limpieza en las calles)

- Codificadas como:

1	Sí
2	No

Tabla 4: Codificación 1 Sí; 2 No

COCINA (Dispone la vivienda de cocina independiente (de 4 m2 o más)
HORNO (Dispone la vivienda de horno)
LAVADORA (Dispone la vivienda de lavadora)
LAVAVAJILLAS (Dispone la vivienda de lavavajillas)
MICROONDAS (Dispone la vivienda de microondas)

SECADORA (Dispone la vivienda de secadora)
VITROINDUC (Dispone la vivienda de vitrocerámica y/o inducción

- Numéricas ordinales:

NASEOS (Numero de cuartos de baño o aseos de la vivienda)
NDORMITO (Numero de dormitorios de la vivienda)
NHIJOS (Número total de hijos (independientemente de si conviven con usted en el hogar o no y de si están vivos actualmente o no)
NOTRASHABIT (Numero de otras estancias de la vivienda excluyendo cocinas, garajes, pasillos, vestíbulos, vestidores, despensas, terrazas abiertas y dependencias utilizadas para fines profesionales)
NPLANTASBAJO (Número de plantas bajo rasante)
NPLANTASSOB (Número de plantas sobre rasante)
NRESI (Número de residentes en la vivienda)
NSALONES (Numero de salones, comedores o cuartos de estar de la vivienda)
NSOTOTRAST (Numero de buhardillas, sótanos o trasteros de la vivienda)
NVEHICULOS (Número de vehículos de los que dispone el hogar)

- Codificadas como escalas ordinales del 0 al 10:

ESTADOEDIF (Estado de conservación del edificio (escala de 0 a 10)
SATISTIEMP (Grado de satisfacción en el tiempo de desplazamiento al trabajo, valorado de 0 a 10 (donde 0 significa totalmente insatisfecho y 10 significa completamente satisfecho).

- Con su propia codificación:

AYUDAEXT (Dispone el hogar de ayudas externas para tareas domésticas y cuidado de menores o mayores dependientes)

1	Sí, dispone de ayudas de familiares, parientes, amigos, vecinos
2	Sí, dispone de ayudas de otros, como servicios sociales o una ONG
3	No dispone de ayudas externas no remuneradas

Tabla 5: Codificación de AYUDAEXT

CALEFAC (Dispone la vivienda de calefacción)

1	No tiene calefacción
2	Sí, colectiva
3	Sí, individual

4	No tiene instalación de calefacción pero sí algún aparato que permite calentar alguna habitación (por ejemplo radiadores eléctricos)

Tabla 6: Codificación de CALEFAC

EC (Estado Civil legal del encuestado)

1	Soltero/a
2	Casado/a en primeras nupcias
3	Casado/a en segundas o más nupcias
4	Viudo/a
5	Separado/a
6	Divorciado/a

Tabla 7: Codificación de EC

ESTUDIOS (Nivel de estudios del encuestado):

1	No sabe leer o escribir
2	Sabe leer y escribir pero fue menos de 5 años a la escuela
3	Educación primaria completa o fue a la escuela al menos 5 años
4	Primera etapa de educación secundaria y similar (EGB, Bachiller elemental, ESO, certificado de Estudios Primarios, certificado de Escolaridad o certificado de Profesionalidad niveles 1 o 2)
5	Segunda etapa de educación secundaria con orientación general (BUP, COU, PREU, Bachiller Superior)
6	Segunda etapa de educación secundaria con orientación profesional (FP grado medio, FPI)
7	Educación postsecundaria no superior (Certificado de Profesionalidad nivel 3, título propio universitario de menos de 2 años que requiere bachillerato)
8	Enseñanzas de formación profesional, artes plásticas y diseño y deportivas de grado superior y equivalentes; títulos propios universitarios que precisan del título de bachiller, de duración igual o superior a 2 años. (FP de grado superior, FPII)
9	Grados universitarios de hasta 240 créditos ECTS, diplomados universitarios, títulos propios universitarios de experto o especialista, y similares
10	Grados universitarios de más de 240 créditos ECTS, licenciados
11	Másteres, especialidades en Ciencias de la Salud por el sistema de residencia y similares
12	Doctorado universitario

Tabla 8: Codificación de ESTUDIOS

INGREHOG (Ingresos mensuales netos del hogar en intervalos)

1	Menos de 500€
2	De 500€ a menos de 1.000€
3	De 1.000€ a menos de 1.500€
4	De 1.500€ a menos de 2.000€
5	De 2.000€ a menos de 2.500€
6	De 2.500€ a menos de 3.000€
7	De 3.000€ a menos de 5.000€
8	De 5.000€ a menos de 7.500€
9	7.500€ o más

Tabla 9: Codificación de INGREHOG

LUGTRAB (Dónde está el lugar de trabajo/estudio)

1	En el propio domicilio
2	En varios municipios (soy comercial, repartidor, taxista…)
3	En el municipio en el que resido
4	En otro municipio de la misma provincia
5	En otro municipio de otra provincia
6	En otro país

Tabla 10: Codificación de LUGTRAB

METROSVI (Metros cuadrados de la vivienda)

1	Hasta 30 m2
2	Entre 31 y 45 m2
3	Entre 46 y 60 m2
4	Entre 61 y 75 m2
5	Entre 76 y 90 m2
6	Entre 91 y 105 m2
7	Entre 106 y 120 m2
8	Entre 121 y 150 m2
9	Entre 151 y 180 m2
10	Más de 180 m2

Tabla 11: Codificación de METROSVI

MTRANSPOR_1 (Primer medio de transporte que utiliza para ir de su casa hasta el lugar de trabajo/estudio con el que cubre más distancia)

1	Coche - particular
2	Coche - de empresa
3	Coche - de terceros - taxi
4	Coche - de terceros - VTC
5	Coche - de terceros - de una compañía sharing
6	Coche - de terceros - vehiculo compartido
7	En moto - particular
8	En moto - de una compañía sharing
9	En bicicleta - particular
10	En bicicleta - de una compañía de sharing
11	En autobús, autocar, minibús - transporte público
12	En autobús, autocar, minibús - servicio de empresa
13	Otros - en metro
14	Otros - en tranvía o metro ligero
15	Otros - andando
16	Otros - en tren
17	Otros - otros medios

Tabla 12: Codificación de MTRANSPOR_1

NDESPLA (Número de desplazamientos totales al día desde su vivienda hasta el lugar de trabajo/estudio)

1	1
2	2
3	3
4	4
5	Más de 4 desplazamientos

Tabla 13: Codificación de NDESPLA

NPLAZASGAR (Número de plazas de aparcamiento del garaje)

1	1
2	2
3	De 3 a 5
4	De 6 a 10
5	De 11 a 20
6	De 21 a 50
7	De 51 a 100
8	De 101 a 150
9	Más de 150

Tabla 14: Codificación de NPLAZASGAR

NVIVIENDP (Número de viviendas que hay en la planta en la que reside)

1	1
2	2
3	3
4	4
5	De 5 a 9
6	10 o más

Tabla 15: Codificación de NVIVIENDP

REGVI (régimen de tenencia de la vivienda)

1	Propia por herencia o donación
2	Propia, por compra, totalmente pagada
3	Propia, por compra, con pagos pendientes (hipotecas)
4	Alquilada
5	Cedida gratis o a bajo precio (por otro hogar, pagada por la empresa…)
6	Otra forma

Tabla 16: Codificación de REGVI

SERVDOMES (Dispone el hogar de servicio doméstico remunerado)

1	Sí, dispone de servicio doméstico interno
2	Sí, dispone de servicio doméstico externo
3	No dispone de servicio doméstico

Tabla 17: Codificación de SERVDOMES

SITLAB (Situación laboral principal durante la última semana)

1	Ocupado/a - A tiempo completo
2	Ocupado/a - A tiempo parcial

3	Estudiante
4	Parado/a - Ha trabajado anteriormente
5	Parado/a - No ha trabajado anteriormente
6	Jubilado/a, prejubilado/a
7	Incapacitado/a permanentemente para trabajar
8	Dedicado/a las tareas del hogar
9	Otro tipo de inactividad

Tabla 18: Codificación de REGVI

TIEMDESPLA (Tiempo que dedica a la suma de todos sus trayectos diarios (ida y vuelta) desde su casa hasta el lugar de trabajo/estudio)

1	Menos de 20 minutos
2	Entre 20 y 39 minutos
3	Entre 40 y 59 minutos
4	Entre 60 y 89 minutos
5	Entre 90 y 119 minutos
6	Entre 2 horas y 2 horas y media
7	Más de 2 horas y media

Tabla 19: Codificación de TIEMDESPLA

TIPOAGUA (Sistema de suministro de agua en su vivienda)

1	Agua corriente por abastecimiento público
2	Agua corriente por abastecimiento privado o particular del edificio
3	No tiene agua corriente

Tabla 20: Codificación de TIPOAGUA

TIPOEDIFVIV (Tipo de edificio en función del número de viviendas)

1	Vivienda unifamiliar (chalet, adosado, pareado…)
2	Edificio con 2 viviendas
3	Edificio de 3 a 9 viviendas
4	Edificio con 10 o más viviendas

Tabla 21: Codificación de TIPOEDIFVIV

5.3 CREACIÓN DE LA BASE DE DATOS FINAL

Para poder realizar el posterior análisis, se han juntado en una misma base de datos las variables independientes (mencionadas en 4.4) y la variable dependiente (4.3).

Se ha realizado la unión de bases por el código del municipio en los casos que es posible y por el código de la provincia en el caso de los municipios de los que no se dispone del dato completo, como ya se ha comentado.

5.4. SELECCIÓN DE LA MUESTRA

Debido al alto número de observaciones del conjunto original y estando descompensadas por tener muchas más de los municipios más grandes que de los pequeños, se ha realizado una selección por muestreo aleatorio simple sin reposición, seleccionando un tamaño muestral de 100 en localidades con más de 50.000 habitantes y 1.000 observaciones por provincia de municipios con un número inferior de habitantes al anteriormente mencionado.

En conclusión, en la muestra se trabajará con 6.510 observaciones de las 50 provincias españolas, Ceuta y Melilla.

6. ANÁLISIS DESCRIPTIVO DE LAS VARIABLES

A continuación, se analizan las características de las variables del presente estudio. En concreto, se analizarán la variable dependiente y las que mayor nivel de asociación tengan con esta debido al alto número de variables en el modelo. Este análisis, al igual que los otros se hará con el software R y la herramienta RStudio.

6.1 VARIABLE DEPENDIENTE

Como se ha comentado en 4.1. la variable dependiente es una variable cualitativa con 4 niveles, los cuales indican el nivel de riesgo de despoblación en la que se encuentra el municipio. El primer nivel (X1) significaría que el municipio está en proceso de despoblación, X2 en riesgo de despoblación, X3 sin riesgo y X4 en sobrepoblación.

La distribución que presenta en el conjunto de datos original es las siguiente:

DESPOBLACIÓN	RIESGO	SIN RIESGO	SOBREPOBLACIÓN
3,183177%	15,970868%	7,946476	72,899479%

Tabla 22: Distribución de la variable dependiente previo muestreo

Como podemos observar, la mayoría de los datos pertenecen a la última categoría (debido a que se ha entrevistado a más personas de grandes municipios que de pequeños), por lo que es necesario un ajuste de la variable, motivo por el cual se aplica el muestreo previamente explicado.

Tras el muestreo, la nueva y definitiva distribución de los datos de la variable dependiente es:

DESPOBLACIÓN	RIESGO	SIN RIESGO	SOBREPOBLACIÓN
10,75269%	36,25192%	11,52074%	41,47465%

Tabla 23: Distribución de la variable dependiente postmuestreo

La cual es mucho más equilibrada.

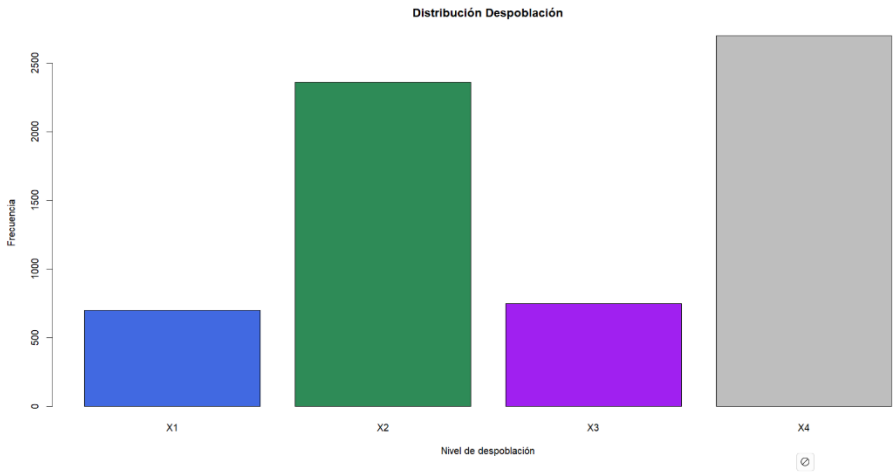

Ilustración 2: Gráfico de la distribución de la variable dependiente

Como podemos observar gráficamente, ahora en lugar de pertenecer la mayoría de los datos a la categoría de sobrepoblación, está más equilibrado entre todas las categorías, perteneciendo un 10% a la categoría de despoblación, un 36% a riesgo de despoblación, un 11% de datos pertenecientes a municipios sin riesgo de despoblación y un 41% de municipios con sobrepoblación.

6.2. ANÁLISIS DE ASOCIACIONES 2 A 2

Para realizar el análisis de correlaciones 2 a 2 de las variables explicativas respecto de la variable dependiente DESPOBLACIÓN se hace uso del Coeficiente de Contingencia y V de Cramer.

Las variables con mayor nivel de asociación con la variable dependiente serían CALEFAC, AIREACON, GASTUBERIA y NPLANTASSOB, como se puede observar en ANEXO 1.

- **AIREACOND**

En cuanto a la tenencia de aire acondicionado en las viviendas, se obtiene un Coeficiente de Contingencia de 0.498 y una V de Cramer de 0.574. Es una relación moderada, pues el Coeficiente de Contingencia no alcanza el 0.5 pero significativa, pues la V de Cramer es superior a 0.3.

- **CALEFAC**

La disposición del tipo de calefacción de la vivienda arroja una asociación del Coeficiente de Contingencia de 0.612 y respecto a la V de Cramer de 0.446, lo que se traduce en una asociación

significativa entre el nivel de despoblación en el que se encuentran las viviendas y el tipo de calefacción de esta.

- **GASTUBERIA**

La disponibilidad del agua por tubería del edificio respecto de la despoblación del municipio en el que se encuentra la vivienda tiene un Coeficiente de Contingencia de 0.483 y una V de Cramer de 0.551, lo que implica una relación moderada entre la variable dependiente y la independiente.

- **NPLANTASSOB**

El número de plantas por encima de la planta baja indica una alta correlación en cuanto al Coeficiente de Contingencia, con un valor de 0.807 y una V de Cramer de 0.29, lo que indica que la asociación tras la normalización no es tan fuerte como inicialmente con el Coeficiente de Contingencia, pero existente.

Como se observa, ninguna variable tiene una asociación muy elevada con la despoblación, la más alta sería en cuanto al Coeficiente de Contingencia sería el número de plantas por encima de la planta baja (con el problema de que tiene una V de Cramer muy baja) y en cuanto a la otra métrica, sería la disponibilidad del gas por tubería de la vivienda.

6.3. VARIABLES EXPLICATIVAS

La distribución de las 4 variables explicativas más asociadas con la variable dependiente es:

- **AIREACOND**

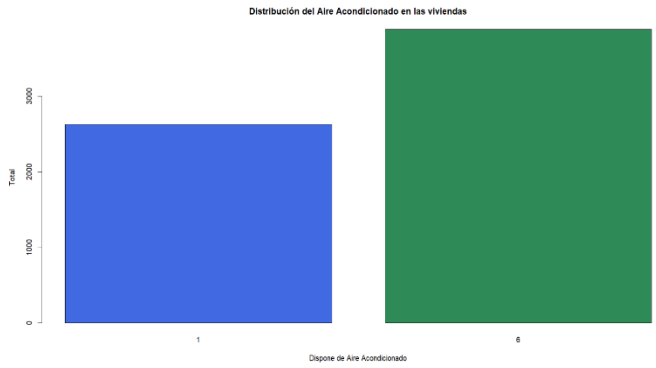

Ilustración 3: Distribución del Aire Acondicionado en las viviendas

Del total de las 6510 observaciones, 3885 no disponen de aire acondicionado, mientras que 2625 sí disponen de éste. Para profundizar en su distribución, calculamos una tabla de contingencia con las proporciones de la variable dependiente y la independiente:

DESPOBLACION\AIREACOND	**1 (SÍ)**	**6 (NO)**
X1 (Despoblación)	0.1471429	0.8528571

X2 (Riesgo)	0.3949153	0.6050847
X3 (Sin riesgo)	0.4053333	0.5946667
X4 (Sobrepoblación)	0.4762963	0.5237037

Tabla 24: Proporción de viviendas con disponibilidad de aire acondicionado y tamaño del municipio

Como podemos ver, a medida que aumenta el tamaño del municipio en el que se encuentra la vivienda, aumenta la proporción de viviendas con aire acondicionado.

- **CALEFAC**

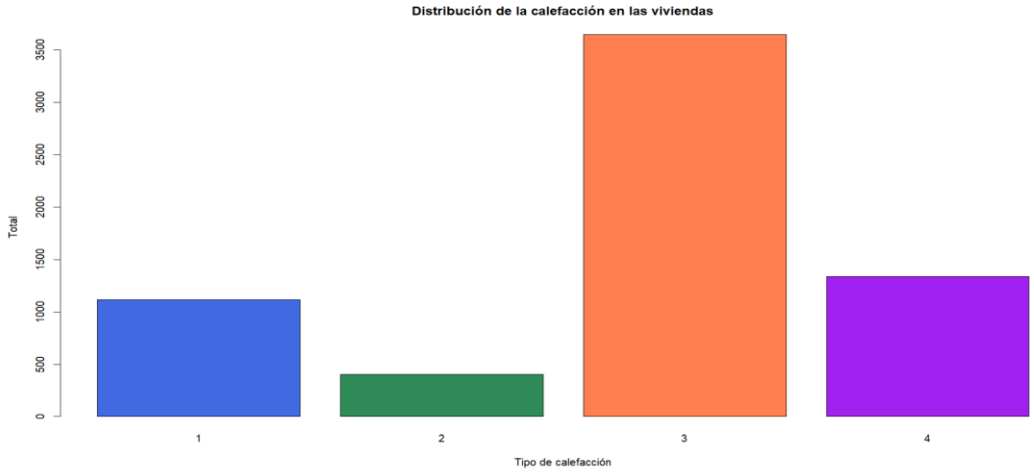

Ilustración 4: Distribución de la calefacción en las viviendas

La mayoría de las viviendas del conjunto de datos, un total de 3649, tienen calefacción individual, seguido de algún dispositivo que caliente alguna habitación con 1338, y por último no tener calefacción con 1118 observaciones y calefacción colectiva con 405.

Observando, al igual que en la anterior variable la tabla de contingencia de las proporciones con la variable dependiente, obtenemos la siguiente tabla representada en porcentajes:

TIPO DE CALEFACCIÓN/ DESPOBLACIÓN	**1 (NO DISPONE)**	**2 (COLECTIVA)**	**3 (INDIVIDUAL)**	**4 (APARATOS)**
X1(Despoblación)	5,14	5,14	77,28	12,42
X2 (Riesgo)	13,81	4,70	61,01	20,46
X3 (Sin riesgo)	19,60	5,20	48,80	26,40
X4(Sobrepoblación)	22,56	81,11	48,22	21,11

Tabla 25: Proporción de viviendas por tipo de aire acondicionado y tamaño del municipio

Se puede ver cómo a medida que aumenta el tamaño del municipio aumenta la proporción de viviendas sin calefacción, mientras que la proporción de viviendas con calefacción individual disminuye.

- **GASTUBERIA**

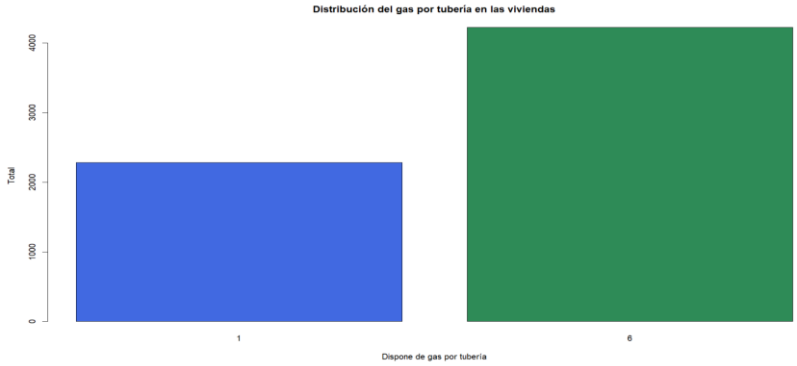

Ilustración 5: Distribución del gas por tubería

De las 6510 observaciones de la muestra, 2285 disponen de gas por tubería mientras que 4225 no disponen de él.

	1 (SÍ)	6 (NO)
X1(Despoblación)	17,71	82,28
X2 (Riesgo)	24,11	75,88
X3 (Sin riesgo)	29,86	70,13
X4(Sobrepoblación)	50,67	49,33

Tabla 26: Proporción de viviendas por disponibilidad de gas por tubería y tamaño del municipio

A medida que aumenta el tamaño del municipio, la disponibilidad del gas por tubería aumenta.

- **NPLANTASSOB**

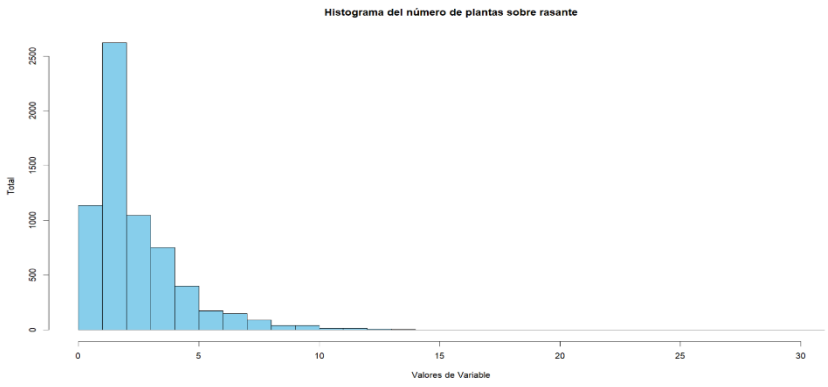

Ilustración 6: Histograma del número de plantas sobre rasante del edificio

Cuantas más plantas sobre rasante menor número de viviendas, siendo mayoritarios los edificios con una planta.

En cuanto a la relación con la variable dependiente, se observa que cuantas más plantas, mayor tamaño del municipio.

7. MODELO DE REGRESIÓN LOGÍSTICA

Como se ha mencionado anteriormente, la regresión logística es una técnica de Machine Learning empleada para crear un modelo de clasificación para una variable dependiente.
En nuestro caso, el objetivo será clasificar las observaciones dentro de las categorías de la variable DESPOBLACIÓN por medio de las variables explicativas mencionadas anteriormente.
El modelo completo por tanto tendría 132 parámetros y 61 variables, el número de parámetros resultante de la creación de variables *dummy* en el caso de variables categóricas, como por ejemplo en el caso de la variable TIPOAGUA, con 3 categorías, que genera 3 parámetros. La forma de construcción de estas variables es la siguiente:
Si un individuo pertenece a la categoría "Agua corriente por abastecimiento público" de TIPOAGUA, entonces aparecerá un 1 multiplicando al parámetro β que le corresponda, multiplicando un 0 a los otros dos parámetros correspondientes a las categorías "Agua corriente por abastecimiento privado o particular del edificio" y "No tiene agua corriente".
Previo al análisis, se realiza una partición de datos en entrenamiento/prueba para la evaluación del modelo. El resultado serían 5.208 en el conjunto de entrenamiento y 1.302 en el de prueba.

7.1. SELECCIÓN DE VARIABLES

Dada la complejidad del modelo, sumado al alto número de observaciones, conviene reducir tanto el número de parámetros como de variables para el análisis. Por ello, realizamos un *Analysis of Deviance* para ver la significatividad de las variables del modelo completo.
Interpretamos como variable significativa en términos generales aquella en la que un cambio en la variable explicativa genera un cambio en la variable respuesta, que en la siguiente salida de R podemos observar en la columna Pr(>Chisq) en los valores inferiores a 0.05, que sería el nivel de significación fijado.
En este caso, al estar hablando de una regresión logística, las hipótesis serían:
H0: $P \leq 0.05$: La asociación es estadísticamente significativa.
H1: $P>0.05$: La asociación no es estadísticamente significativa

7.1.1 MODELO COMPLETO

Response: DESPOBLACION

	LR Chisq	Df	Pr(>Chisq)
NRESI	1.163	1	0.2807888
REGVI	8.116	5	0.1499766
CALEFAC	110.570	3	< 2.2e-16 ***
TIPOAGUA	10.281	2	0.0058558 **
AIREACOND	16.003	1	6.325e-05 ***
AISLAM	1.156	1	0.2822263
ADAPTADA	0.195	1	0.6584887
INTERNET	8.356	1	0.0038446 **
COCINA	0.080	1	0.7777010
LAVAVAJILLAS	5.481	1	0.0192230 *
LAVADORA	9.048	1	0.0026296 **
SECADORA	10.411	1	0.0012526 **
HORNO	1.886	1	0.1696911
MICROONDAS	2.155	1	0.1421012
VITROINDUC	-0.002	1	1.0000000
NASEOS	10.651	1	0.0011002 **
NSALONES	0.991	1	0.3195033
NDORMITO	8.002	1	0.0046723 **
NSOTOTRAST	0.345	1	0.5569359
NOTRASHABIT	0.434	1	0.5101441
METROSVI	36.666	9	3.019e-05 ***
NVEHICULOS	2.738	1	0.0979883 .
SERVDOMES	3.771	2	0.1517473
AYUDAEXT	2.752	2	0.2526129
RUIDOS	0.000	1	1.0000000
CONTAMIN	2.105	1	0.1468202
SUCIO	5.302	1	0.0212957 *
MALCOMUNIC	3.742	1	0.0530552 .
POCOVERDE	0.046	1	0.8306875
DELINCUENCIA	4.187	1	0.0407251 *
MOLESTURIST	0.005	1	0.9441308

COLEGIO	0.006	1	0.9391395
CSALUD	2.581	1	0.1081287
SUPER	7.885	1	0.0049833 **
FARMACIA	0.001	1	0.9817616
BARES	2.407	1	0.1207776
SEGUNRESI	0.189	1	0.6638089
INGREHOG	48.226	8	8.942e-08 ***
TIPOEDIFVIV	4.181	3	0.2425423
NPLANTASBAJO	1.166	1	0.2802894
NPLANTASSOB	14.642	1	0.0001300 ***
ESTADOEDIF	1.007	1	0.3155590
GARAJE	0.002	1	0.9636161
NPLAZASGAR	19.140	9	0.0240322 *
GASTUBERIA	21.978	1	2.758e-06 ***
AGUACALCENT	1.370	1	0.2417417
EVACUAGUARES	0.984	1	0.3212084
ENERENOV	0.058	1	0.8097452
NVIVIENDP	3.766	5	0.5835719
EC	3.676	5	0.5968681
ESTUDIOS	26.064	11	0.0063499 **
SITLAB	0.696	8	0.9995369
FLEXI	0.065	2	0.9678185
LUGTRAB	47.955	6	1.207e-08 ***
NDESPLA	23.616	5	0.0002573 ***
MTRANSPOR_1	28.763	15	0.0172506 *
TIEMDESPLA	15.814	7	0.0268683 *
SATISTIEMP	1.588	1	0.2076052
SMART	0.017	1	0.8951469
NHIJOS	1.384	1	0.2394911
PAREJA	0.048	1	0.8270670

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tabla 27: Analysis of Deviance modelo completo

Las variables más significativas en la regresión son, por orden de significación, INGREHOG, AIREACOND, METROSVI, GASTUBERIA y CALEFAC, tres de ellas como pudimos ver estaban altamente relacionadas en el análisis de asociaciones con la variable dependiente.

Basándose en estos resultados, se deduce un nuevo modelo solo con las variables que son significativas en este modelo completo, que será mucho más simple y por lo tanto óptimo de cara a la predicción de la variable dependiente.

Las variables que se han conservar en el modelo por su significación al 0.05, por tanto, serán:

CALEF, TIPOAGUA, AIREACOND, INTERNET, LAVAVAJILLAS, LAVADORA, SECADORA, NASEOS, NDORMITO, METROSVI, SUCIO, DELINCUENCIA, SUPER, INGREHOG, NPLANTASSOB, NPLAZASGAR, GASTUBERIA, ESTUDIOS, LUGTRAB, NDESPLA, MTRANSPOR_1, TIEMDESPLA.

7.1.2. MODELO REDUCIDO

Por motivos de multicolinealidad, algunas variables significativas han de ser eliminadas, que son NDESPLA, MTRANSPOR_1 y TIEMDESPLA. Esto es debido a que, al intentar crear el nuevo modelo en R, aparece el siguiente mensaje:

```
Warning: design appears to be rank-deficient, so dropping some coefs
```

Lo que indica problemas de multicolinealidad, y tras probar qué variables son las que generaban el problema se ha llegado a esta conclusión.

Además, por interés en el estudio, se añaden al modelo BARES y COLEGIO a pesar de no ser significativas en este primer análisis. Si siguieran sin serlo en el nuevo modelo, no se incluirían definitivamente, pero sino sí se incluyen en el modelo final. También se incluye MALCOMUNIC, ya que esté muy levemente por encima del límite de significación.

Así, tras la creación del nuevo modelo, la nueva salida del análisis de tipo II es:

```
Analysis of Deviance Table (Type II tests)

Response: DESPOBLACION
```

LR	Chisq	Df	Pr(>Chisq)
CALEFAC	220.894	3	< 2.2e-16 ***
TIPOAGUA	27.790	2	9.237e-07 ***
AIREACOND	21.788	1	3.045e-06 ***
INTERNET	49.229	1	2.277e-12 ***
NDORMITO	25.617	1	4.164e-07 ***
MALCOMUNIC	12.453	1	0.0004174 ***
COLEGIO	12.602	1	0.0003854 ***
BARES	12.949	1	0.0003201 ***
SUPER	39.187	1	3.851e-10 ***
INGREHOG	99.941	8	< 2.2e-16 ***
NPLANTASSOB	160.669	1	< 2.2e-16 ***
GASTUBERIA	61.626	1	4.153e-15 ***
LUGTRAB	50.585	6	3.587e-09 ***
METROSVI	36.399	9	3.367e-05 ***
DELINCUENCIA	15.313	1	9.110e-05 ***
LAVAVAJILLAS	7.629	1	0.0057442 **
LAVADORA	8.826	1	0.0029703 **
SECADORA	24.374	1	7.934e-07 ***
NPLAZASGAR	56.488	9	6.333e-09 ***

ESTUDIOS	29.402	11	0.0019658 **
NASEOS	10.751	1	0.0010423 **
SUCIO	18.057	1	2.144e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tabla 28: Analysis of Deviance modelo reducido

Ahora todos los parámetros son significativos, por lo que lo tomamos como modelo final.

Tras esta reducción, hemos pasado de un modelo inicial con 132 parámetros y 60 variables a un modelo con 22 variables y 63 parámetros.

7.2. EVALUACIÓN DEL MODELO

7.2.1 PRUEBAS INDIVIDUALES SOBRE LOS PARÁMETROS E INTERPRETACIÓN.

A continuación, se analizan los parámetros con la función *stargazer* de R, que genera una tabla resumen con la información de los parámetros del modelo creado.

```
==========================================
                  Dependent variable:
              ---------------------------
                       DESPOBLACION
------------------------------------------
CALEFAC2                -0.780***
                         (0.148)

CALEFAC3                -1.202***
                         (0.085)

CALEFAC4                -0.542***
                         (0.092)

TIPOAGUA2               0.647***
                         (0.130)

TIPOAGUA3                1.546
                         (0.996)

AIREACOND6              -0.286***
                         (0.061)

INTERNET6               -0.539***
                         (0.077)

NDORMITO                -0.173***
                         (0.034)

MALCOMUNIC6             0.261***
                         (0.074)

COLEGIO6                -0.307***
                         (0.086)

BARES6                  0.368***
                         (0.103)

SUPER6                  -0.537***
```

	(0.086)
INGREHOG2	-0.102
	(0.165)
INGREHOG3	0.181
	(0.164)
INGREHOG4	0.318*
	(0.169)
INGREHOG5	0.468***
	(0.175)
INGREHOG6	0.724***
	(0.186)
INGREHOG7	0.715***
	(0.184)
INGREHOG8	1.496***
	(0.261)
INGREHOG9	0.326
	(0.283)
NPLANTASSOB	0.213***
	(0.018)
GASTUBERIA6	-0.550***
	(0.070)
LUGTRAB1	-0.281*
	(0.149)
LUGTRAB2	-0.002
	(0.147)
LUGTRAB3	-0.227***
	(0.080)
LUGTRAB4	0.012
	(0.083)
LUGTRAB5	-0.910***
	(0.146)
LUGTRAB6	0.600
	(0.479)
DELINCUENCIA6	-0.379***
	(0.098)
LAVAVAJILLAS2	0.182***
	(0.066)
LAVADORA2	-0.820***
	(0.277)
SECADORA2	-0.328***
	(0.067)
NPLAZASGAR1	-0.300***
	(0.078)
NPLAZASGAR2	-0.225**
	(0.096)
NPLAZASGAR3	-0.226*

	(0.130)
NPLAZASGAR4	-0.095
	(0.170)
NPLAZASGAR5	-0.115
	(0.136)
NPLAZASGAR6	0.365***
	(0.132)
NPLAZASGAR7	0.534***
	(0.192)
NPLAZASGAR8	0.318
	(0.351)
NPLAZASGAR9	1.291***
	(0.349)
NASEOS	0.157***
	(0.048)
SUCIO6	-0.331***
	(0.078)
ESTUDIOS10	0.229
	(0.301)
ESTUDIOS11	0.556
	(0.396)
ESTUDIOS12	0.063
	(0.472)
ESTUDIOS2	-0.189
	(0.291)
ESTUDIOS3	-0.247
	(0.283)
ESTUDIOS4	-0.195
	(0.281)
ESTUDIOS5	-0.122
	(0.293)
ESTUDIOS6	-0.133
	(0.294)
ESTUDIOS7	0.586
	(0.431)
ESTUDIOS8	0.108
	(0.298)
ESTUDIOS9	0.005
	(0.298)
METROSVI10	-0.309
	(0.392)
METROSVI2	0.045
	(0.440)
METROSVI3	-0.076
	(0.390)
METROSVI4	-0.059

```
                          (0.382)

METROSVI5                -0.156
                          (0.378)

METROSVI6                -0.534
                          (0.380)

METROSVI7                -0.349
                          (0.384)

METROSVI8                -0.569
                          (0.387)

METROSVI9                -0.321
                          (0.393)

---------------------------------------
Observations               5,208
=======================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Tabla 29: Resumen parámetros modelo reducido

En esta tabla podemos observar el nivel de significación de los parámetros, junto con los coeficientes y el error estándar de cada uno entre paréntesis

Como podemos ver hay variables con ninguno de sus parámetros son significativos, lo que es una contradicción estadística dado que las variables anteriormente sí lo eran. Esto se debe, de nuevo, a problemas de multicolinealidad por lo que eliminamos estas variables del modelo. Serían: METROSVI y ESTUDIOS.

Tras ello volvemos a analizar la significatividad de las variables y los parámetros.

Analysis of Deviance Table (Type II tests)

Response: DESPOBLACION

	LR Chisq	Df	Pr(>Chisq)
CALEFAC	224.527	3	< 2.2e-16 ***
TIPOAGUA	29.756	2	3.455e-07 ***
AIREACOND	18.553	1	1.652e-05 ***
INTERNET	56.886	1	4.619e-14 ***
NDORMITO	46.180	1	1.079e-11 ***
MALCOMUNIC	10.910	1	0.0009562 ***
COLEGIO	13.129	1	0.0002907 ***
BARES	13.949	1	0.0001878 ***
SUPER	43.890	1	3.474e-11 ***

INGREHOG	113.420	8	< 2.2e-16 ***
NPLANTASSOB	173.075	1	< 2.2e-16 ***
GASTUBERIA	72.315	1	< 2.2e-16 ***
LUGTRAB	44.341	6	6.325e-08 ***
DELINCUENCIA	16.587	1	4.648e-05 ***
LAVAVAJILLAS	8.964	1	0.0027529 **
LAVADORA	8.366	1	0.0038230 **
SECADORA	22.537	1	2.061e-06 ***
NPLAZASGAR	71.561	9	7.521e-12 ***
NASEOS	10.323	1	0.0013136 **
SUCIO	20.104	1	7.334e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tabla 30: Anlysis of Deviance modelo final

Todas las variables continúan siendo significativas.

Como se puede comprobar en el ANEXO 2, no existen variables sin parámetros significativos y ahora podemos interpretar los parámetros a través de los ODDS-RATIO, que nos permite cuantificar el efecto de las
variables independientes sobre la dependiente. La tabla con el resumen de los valores de los parámetros se encuentren en ANEXO 3.

- **CALEFAC**

Tener calefacción o algún sistema electrónico que permita calentar una casa reduce las probabilidades de que la vivienda esté en un municipio con alto número de habitantes.
Concretamente, tener calefacción colectiva lo reduce en un 46%, tener calefacción individual en un 70% y no tener un sistema, pero sí algún aparato, lo reduce en un 42%.

- **TIPOAGUA**

Disponer de agua corriente por abastecimiento privado o particular del edificio incrementa la probabilidad de pertenecer a un municipio más grande en un 95%.

- **AIREACOND**

No tener aire acondicionado reduce en un 33% la probabilidad de pertenecer a un municipio con sobrepoblación, lo que sería lo mismo que en caso de tenerlo, hay un 77% más de probabilidad de pertenecer a un gran municipio.

- **NDORMITO**

Aumentar el número de dormitorios reduce en un 20% la probabilidad de pertenecer a grandes municipios.

- **MALCOMUNIC**

Que la vivienda no esté mal comunicada aumenta en un 27% las probabilidades de vivir en municipios de mayor tamaño en detrimento de municipios con menos habitantes.

- **COLEGIO**

No disponer de colegio cercano a la vivienda reduce en un 27% la probabilidad de que se encuentre en una zona con más población.

- **BARES**

No disponer de bares cercanos a la vivienda aumenta en un 46% la probabilidad de vivir en un área con más población.

- **SUPER**

No disponer de un supermercado cercano a la vivienda reduce en un 46% la probabilidad de vivir en un municipio con alto nivel de población.

- **INGREHOG**

Tener ingresos de menos de 500€ reducen en un 10% la pertenencia aun gran municipio mientras que ingresos de entre 500€, 1000€, 1500€, 2000€, 2500€, 3000€ y 5000€ aumentan las probabilidades un 20%, 35% 60%, 108%, 113%, 385% respectivamente.

- **NPLANTASSOB**

Por cada planta adicional que se sume al bajo, aumenta en un 24% vivir en un municipio más grande.

- **GASTUBERIA**

No disponer de gas por tubería reduce en un 45% la probabilidad de que la vivienda no pertenezca a un gran municipio.

- **LUGTRAB**

Trabajar en el propio domicilio reduce en un 17% las probabilidades de vivir en un gran municipio, al igual que trabajar en el mismo municipio y trabajar en otro municipio de otra provincia (reduciendo en un 55% la probabilidad de trabajar en un gran municipio)-

Al contrario, trabajar en varios municipios, en otro municipio de la provincia o en otro país aumentan en un 0.8%, 0.9% y un 110% respectivamente.

- **DELINCUENCIA**

Que no haya delincuencia en el municipio reduce en un 33% las probabilidades de pertenecer a un gran municipio.

- **LAVAVAJILLAS, LAVADORA, SECADORA**

No disponer de lavavajillas aumenta en un 21% las probabilidades de pertenecer a municipios de mayor tamaño, mientras que no tener lavadora reduce en un 55% las posibilidades y no tener secadora lo reducen un 27%.

- **NPLAZASGAR**

Tener 20 plazas de garaje o menos en el garaje reduce las probabilidades de que la vivienda esté ubicada en un gran municipio, concretamente, para 1 plaza se reduce en un 30%, para 2 y de 3 a 5 en un 25%, de 6 a 10 en un 9% y de 11 a 20 en un 12%.

Aumenta la probabilidad de que la variable DESPOBLACION tome mayores valores de 21 a 50 plazas de garaje en un 45%, de 51 a 100 en un 75%, de 101 a 150 en un 39% y más de 150 plazas en un 279%.

- **NASEOS**

Un aumento unitario en el número de aseos de la vivienda aumenta en un 15% la probabilidad de que aumente el tamaño del municipio.

- **SUCIO**

Tener limpieza en las calles aumenta en un 30% la probabilidad de que un municipio pertenezca a municipios de mayor tamaño.

7.2.2 MATRIZ DE CONFUSIÓN

La matriz de confusión resultante, obtenida con *confusionMatrix* de la librería *caret*, es:

Confusion Matrix and Statistics

	X1	X2	X3	X4
X1	51	44	0	3
X2	436	1261	285	541
X3	0	0	0	0
X4	73	583	315	1616

Tabla 32: Matriz de confusión modelo final

En el eje vertical se representan los casos reales de las categorías y en el eje horizontal las categorías predichas.

Como se observa, de las 560 observaciones propias de viviendas en municipios en despoblación (X1) se clasificaron correctamente 51, mientras que 44 fueron clasificadas como X2 y 3 como X4.

De las 1.888 observaciones propias de municipios en riesgo de despoblación, se clasificaron correctamente 1261, mientras que 436 fueron clasificadas como X1, como X4, 541 y como X3 fueron 285.

En cuanto a los municipios sin problemas de despoblación, el modelo no fue capaz de clasificar ninguna observación .

Por último, refiriéndonos a X4 como municipios con sobrepoblación observamos que se clasificaron correctamente 1.616 y erróneamente 73 en X1, 583 en X2 y 315 en X3.

7.2.3 SENSIBILIDAD Y ESPECIFICIDAD

A partir de la misma función del apartado anterior obtenemos la sensibilidad y especificidad de cada una de las categorías de la variable DESPOBLACION:

```
Statistics by Class:
```

	Class: X1	Class: X2	Class: X3	Class: X4
Sensitivity	0.091071	0.6679	0.0000	0.7481
Specificity	0.989888	0.6199	1.0000	0.6814

Tabla 33: Sensibilidad y especificidad del modelo final

- **Sensibilidad**

Para la clase X1, la sensibilidad es de aproximadamente 0,0911. Esto significa que el modelo es capaz de identificar correctamente el 9,11% de las instancias reales de la clase X1. En otras palabras, el modelo tiene un bajo rendimiento en la detección de verdaderos positivos para la clase X1.

Para la clase X2, la sensibilidad es de aproximadamente 0,6679. Esto indica que el modelo es bastante bueno para detectar verdaderos positivos en la clase X2, con una tasa de acierto del 66,79%.

Para la clase X3, la sensibilidad es de 0,000. Esto sugiere que el modelo no es capaz de detectar correctamente ningún verdadero positivo para la clase X3. La sensibilidad es nula en este caso.

Para la clase X4, la sensibilidad es de aproximadamente 0,7481. Esto significa que el modelo tiene un buen desempeño en la identificación de verdaderos positivos para la clase X4, con una tasa de acierto del 74,81%.

- **Especificidad**

Para la clase X1, la especificidad es de aproximadamente 0,9899. Esto indica que el modelo es muy bueno en la identificación de verdaderos negativos para la clase X1, con una tasa de acierto del 98,99%.

Para la clase X2, la especificidad es de aproximadamente 0,6199. Esto sugiere que el modelo tiene un rendimiento moderado en la identificación de verdaderos negativos en la clase X2, con una tasa de acierto del 61,99%.

Para la clase X3, la especificidad es de 1,0000, lo que significa que el modelo es perfecto en la identificación de verdaderos negativos para la clase X3, con una tasa de acierto del 100%.

Para la clase X4, la especificidad es de aproximadamente 0,6814. Esto indica que el modelo tiene un rendimiento moderado en la identificación de verdaderos negativos en la clase X4, con una tasa de acierto del 68,14%.

7.2.4 PRUEBA DE CONCORDANCIA

Como se ha mencionado anteriormente, la prueba de concordancia óptima para el conjunto de datos es el índice Kappa ponderado, pues tiene en cuenta el carácter ordinal de los datos. Esto se refiere a que no es el mismo error clasificar una vivienda perteneciente a un municipio en despoblación (X1) como perteneciente a un municipio en sobrepoblación (X4), a clasificar un domicilio en un municipio en despoblación (X1) en riesgo de despoblación (X2).
La salida del software, de la función *Kappa* de la librería *vcd* sería:

	value	**ASE**	**z**	**Pr(>|z|)**
Unweighted	0.2894	0.01936	14.95	1.580e-50
Weighted	0.3903	0.02089	18.68	6.919e-78

TABLA 34: Valores y significación de Kappa clásico y ponderado

Podemos observar tanto el índice Kappa clásico (“Unweighted”) como el ajustado (“Weighted”). En el caso del ajustado, un fallo en un nivel de la variable se considera dos

tercios de acierto, un fallo en dos niveles de la variable se considera un tercio de acierto y un fallo en 3 niveles se considera un fallo completo.
Que el contraste entre ambos índices sea significativamente distinto de cero implica que la clasificación del modelo es mejor que la que se conseguiría a raíz del azar.

Podemos ver que el valor de Kappa ajustado es mayor, pues fijándonos en la matriz de confusión la mayoría de los fallos se ubican en las categorías próximas a la del acierto.

Por tanto, podemos concluir que el modelo tiene una calidad justa atendiendo al criterio de calidad del modelo del índice Kappa, muy próximo a una calidad moderada y notablemente superior si solamente nos fijáramos en el índice Kappa clásico.

7.2.5 PRECISIÓN DEL MODELO

Atendiendo al criterio de precisión del modelo, obtenemos que el modelo es capaz de clasificar correctamente un 56,1% de las observaciones, obtenido en la misma salida que los resultados anteriores, lo que podemos considerar como un nivel de precisión aceptable ya que es capaz de clasificar más de la mitad de las observaciones en su categoría. Los resultados proporcionados por el software se encuentran en ANEXO 4.

7.2.6 PRUEBA DE BONDAD DE AJUSTE

Para analizar la bondad de ajuste del modelo de regresión logística primero se realiza un contraste con el modelo nulo, obteniéndose un p-valor de 0 lo que indica que existen diferencias significativas entre el modelo final y el modelo constante, por lo que se puede afirmar que el modelo si tiene una capacidad clasificatoria en sí mismo.

Cuando hablamos de regresión lineal el coeficiente de correlación, r y el de determinación, R^2, son medidas útiles para saber cómo de bien se ajusta el modelo a los datos. En regresión logística podemos calcular una medida análoga, conocida como R-statistic o PseudoR.

En cuanto a la métrica de la R^2 de McFadden, se obtiene un valor de 0.14, lo que implica que el modelo creado es capaz de explicar un 14% de la varianza en comparación con el modelo nulo. Este valor es bastante bajo, pues la mayor parte de la varianza de la regresión queda sin explicar por el modelo, pero en términos predictivos continúa siendo mejor que el modelo nulo lo que ya implica una mejora.

7.2.7 ESTABILIDAD DEL MODELO

Para evaluar la estabilidad del modelo recurrimos a la validación cruzada con los conjuntos de entrenamiento y prueba, calculando las métricas de accuracy y Kappa ponderado para el conjunto de prueba (para el conjunto de entrenamiento ya han sido calculados en 7.2 y 7.3).
Se obtiene una precisión de 0,56144 en el conjunto de prueba y un Kappa ponderado de 0,2893, diferencias prácticamente nulas respecto del conjunto de entrenamiento lo que nos indica que el modelo está balanceado y no está sobreajustado a los datos de entrenamiento y por tanto, si se calcularan otras submuestras las métricas del modelo no distarían notablemente de esta.

8. CONCLUSIONES

El principal objetivo de este trabajo era la creación de un modelo de Machine Learning que permitiera clasificar, sin información geográfica del municipio, las viviendas de un conjunto de datos en función de las características socioeconómicas de dichos domicilios y de los residentes en cada uno de ellos, para poder determinar si pertenecen a municipios afectados por el problema de la despoblación o no en todo el territorio español. Tras ello, se exponen las conclusiones del estudio,

Se ha observado que las variables más asociadas con la despoblación son las referidas a las características de la vivienda, específicamente los recursos de los que dispone pues las variables que mayor asociación presentan son las relacionadas con la disponibilidad de gas por tubería, el tipo de calefacción del domicilio y la disponibilidad del aire acondicionado.
Además, el tamaño de los edificios también se ve asociado al número de habitantes de los municipios, pues a mayor altura menor densidad poblacional se encuentra.
Por ello, a grandes rasgos podemos decir en cuanto a estas cuatro características que las viviendas de los municipios con despoblación tienden a ser más bajas en altura, con menos proporción de ellas con aire acondicionado y de gas por tubería, pero con más calefacciones individuales que el resto de las localidades.

Todas ellas son significativas también no solo en términos asociativos con la densidad de población, sino que también lo son en términos regresivos en la muestra. Más concretamente, hablando de estas variables anteriormente mencionadas podemos decir con mayor precisión que la posesión de aire acondicionado aumenta en un 33% la probabilidad de pertenecer a un municipio pequeño, tener calefacción individual lo reduce en un 70% y no disponer de gas por tubería lo reduce en un 45%. En cuanto a la altura del edificio, aumenta un 24% la probabilidad

de vivir en un municipio que no esté en estado de despoblación o más grande por cada planta que añadimos al bajo, al igual que aumentar unitariamente el número de aseos de la vivienda, que lo incrementa en un 15%. Aumentar el número de dormitorios del domicilio tiene el efecto contrario, reduce la probabilidad de que el municipio sea grande al igual que tener más de 20 plazas de garaje.

En cuanto a la disponibilidad del agua del edificio, que no sea de abastecimiento público aumenta la probabilidad de que el municipio sea más grande.

Tener lavavajillas aumenta las probabilidades de que el municipio sea más pequeño, mientras que tener lavadora y secadora lo reducen.

Refiriéndonos a las características del municipio, aumenta el tamaño el hecho de tener colegio y supermercado próximo a la vivienda, pero al contrario ocurre con tener bares próximos a las viviendas. Una buena limpieza en las calles supone una mayor probabilidad de mayor tamaño del municipio al igual que disponer de buenas comunicaciones, mientras que, si existe delincuencia, se reduce.

En cuanto a las características de las familias existe una relación directa entre los ingresos y el tamaño del municipio en el que se encuentra la vivienda, pues a mayores ingresos aumenta la probabilidad de que se resida en municipios de mayor tamaño, así como que el lugar de trabajo sea en varios municipios, en otro municipio de la provincia o en otro país.

En resumen, en este estudio las características significativas que hacen que las viviendas pertenezcan a municipios con menor población son que disponga de algún tipo de calefacción y gas por tubería, que no disponga de aire acondicionado ni internet, ni colegio cercano a la vivienda, ni supermercado. Tener lavavajillas también contribuye a que el municipio no sea de grandes dimensiones. Los municipios pequeños tienden a tener menos delincuencia y también menos limpieza en las calles y a estar peor comunicados.

A mayor número de dormitorios mayor probabilidad de que el municipio sea pequeño, al contrario que con el número de plazas de garaje y número de aseos y el número de plantas del edificio y los ingresos son bajos.

Por otro lado, la información obtenida sobre la evaluación del modelo es:

El modelo es capaz de clasificar muy bien los municipios con sobrepoblación y en riesgo de despoblación, pero no es tan bueno clasificando municipios en despoblación y no es capaz de detectar municipios sin riesgo (motivado por la descompensación que tenía en la distribución de la variable dependiente). El índice de Kappa ajustado es de 0.39, un valor justo (casi

moderado) que indica que el nivel no es bajo, pero aun así no es un valor del todo aceptable. El modelo clasifica correctamente un 56.1% de las observaciones, que ya sería más de la mitad por lo que es un valor aceptable (hay que tener en cuenta que el modelo no es capaz de clasificar X3) y se rechaza que la capacidad predictiva sea la misma que la del modelo nulo, mejorando la precisión respecto a este en un 14%. Por otro lado, el modelo está muy equilibrado, pues la diferencia entre el conjunto de prueba y entrenamiento es prácticamente nula, por lo que no tiende al sobreajuste.

8.1. POSIBLES MEJORAS DEL ESTUDIO

Dado que el modelo no es del todo bueno, sino más bien de una calidad moderada, se proponen en este apartado posibles mejoras que podrían darse para la continuación y mejora del modelo de aprendizaje automático presentado.

En un principio, se podría tratar de seleccionar mayores observaciones de las categorías X1 (despoblación) y X3 (sin riesgo de despoblación) para que la variable respuesta estuviera más balanceada, y así tratar de mejorar la sensibilidad de ambas, sobre todo de X3, categoría en la que es nula.

También, se podría hacer uso del boostrapping, que en lugar de coger una única muestra como se ha realizado en este análisis realiza un remuestreo (no selecciona una única muestra, sino varias) y crea las métricas del modelo a partir de todos en conjunto, lo que produce una mayor exactitud (sobre todo en este caso, que la muestra es muy pequeña respecto del total de observaciones disponibles en el caso original).

El modelo podría ser mucho más exacto si se dispusieran de todos los códigos de municipios del INE, pues en este caso se ha hecho el análisis a partir de la media de los municipios con menos de 50.000 habitantes en cada provincia, es decir, con una estimación de la densidad poblacional, pero si se dispusiera de la densidad real de cada uno (que sí que existe ya que está disponible en el conjunto de datos del Instituto Geográfico Nacional) el aprendizaje automático sería mucho más preciso.

También, se podría estimar la población con una regresión metiendo las diferentes variables como independientes con el objetivo de comprobar si esta interpolación funciona mejor que la media calculada en el presente estudio para los municipios con menos de 50.000 habitantes. Para ello, habría que suponer que las estimaciones fueran válidas para rangos de habitantes menores que 5.000 habitantes.

9. BIBLIOGRAFÍA

Bortz J., Lienert G. A. & Boehnke K. (2016). *Estadística descriptiva, vol. 18, pp 45-49.*

CALVIÑO, A. (2022). *APUNTES DE SEGMENTACIÓN Y TRATAMIENTO DE ENCUESTAS.* MADRID. https://cvmdp.ucm.es/moodle/pluginfile.php/2706470/mod_resource/content/3/Tema%202_ordinal.pdf

COMISIÓN EUROPEA. (2013). *Directrices sobre las ayudas estatales de finalidad regional para 2014-2020.* https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2013:209:0001:0045:ES:PDF

Flores, Nelly (2022). *Cross validation: qué es y su relación con el Machine Learning*

Isea, R., Ojeda, V., Fernandez, J., Gutierrez, A., & Salazar, V. (2018). *COEFICIENTE V DE CRAMER.* Caracas. https://mariafatimadossantosestadistica1.files.wordpress.com/2018/06/coeficientes-v-de-cramer-y-c-de-pearson.pdf

Hlavac, Marck. *Stargazer: beautiful LATEX, HTML and ASCII tables from R statistical output.* https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

Hastie, T. J. and Pregibon, D. (1992) *Generalized linear models.* Chapter 6 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hu, B., Shao, J., & Palta, M. (2006). *PSEUDO-r 2 in logistic regression model.* Statistica Sinica*,* 16*(3), 847–860.* http://www.jstor.org/stable/24307577

Meyer, David (2023). *Kappa: Cohen's Kappa and Weighted Kappa. https://rdrr.io/cran/vcd/man/Kappa.html*

Mehryar Mohri, A. R. (2018). *Foundations of Machine Learning.* MIT Press.

Petrie, Adam. *Confusion matrix for logistic regression models*. https://search.r-project.org/CRAN/refmans/regclass/html/confusion_matrix.html

PRESIDENTE DEL COMITÉ EUROPEO DE LAS REGIONES. (10 DE DICIEMBRE DE 2020). *Dictamen del Comité Europeo de las Regiones — Estrategia de la UE para la recuperación de las zonas.* BRUSELAS. https://eur-lex.europa.eu/legalcontent/ES/TXT/PDF/?uri=CELEX:52020IR1066&from=EN

SUSI GARCÍA, M. R. S. (2022). *APUNTES DE ESTADÍSTICA APLICADA A LAS CIENCIAS DE LA SALUD. MADRID.* https://cvmdp.ucm.es/moodle/pluginfile.php/2834580/mod_resource/content/4/Tema1_I_Pruebasdiagnostico_Estudiantes.pdf

VICEPRESIDENCIA CUARTA Y MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO. (2020). *EL RETO DEMOGRÁFICO Y LA DESPOBLACIÓN DE ESPAÑA EN CIFRAS.* SECRETARÍA GENERAL PARA EL RETO DEMOGRÁFICO. Obtenido de https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/280220-despoblacion-en-cifras.pdf

Zofel P. (2017) *Coeficiente de contingencia, vol 25, pp 10.*

10. ANEXOS

```
Contingency Coeff.: 0.612
Cramer's V        : 0.446
[1] "CALEFAC"

Contingency Coeff.: 0.498
Cramer's V        : 0.574
[1] "AIREACOND"

Contingency Coeff.: 0.807
Cramer's V        : 0.291
[1] "NPLANTASSOB"

Contingency Coeff.: 0.483
Cramer's V        : 0.551
[1] "GASTUBERIA"
```

ANEXO 1: Asociación de variables independientes más relevantes con dependiente.

```
=========================================
Dependent variable:
---------------------------
DESPOBLACION
-----------------------------------------
CALEFAC2                0.463***
(0.147)

CALEFAC3                0.300***
(0.085)

CALEFAC4                0.583***
(0.091)

TIPOAGUA2               1.957***
(0.129)

TIPOAGUA3                 4.281
(1.030)

AIREACOND6              0.770***
(0.061)

INTERNET6               0.569***
(0.075)

NDORMITO                0.805***
(0.032)

MALCOMUNIC6             1.275***
(0.073)

COLEGIO6                0.733***
(0.086)
```

BARES6	1.462***
	(0.102)
SUPER6	0.568***
	(0.085)
INGREHOG2	0.903
	(0.164)
INGREHOG3	1.204
	(0.162)
INGREHOG4	1.360*
	(0.168)
INGREHOG5	1.608***
	(0.174)
INGREHOG6	2.088***
	(0.184)
INGREHOG7	2.134***
	(0.181)
INGREHOG8	4.850***
	(0.258)
INGREHOG9	1.513
	(0.279)
NPLANTASSOB	1.244***
	(0.018)
GASTUBERIA6	0.554***
	(0.070)
LUGTRAB1	0.830
	(0.146)
LUGTRAB2	1.009
	(0.145)
LUGTRAB3	0.838**
	(0.078)
LUGTRAB4	1.098
	(0.079)
LUGTRAB5	0.455***
	(0.142)
LUGTRAB6	2.100
	(0.482)
DELINCUENCIA6	0.675***
	(0.097)
LAVAVAJILLAS2	1.216***
	(0.065)
LAVADORA2	0.450***
	(0.277)
SECADORA2	0.730***
	(0.066)
NPLAZASGAR1	0.701***
	(0.077)

NPLAZASGAR2	0.754***
	(0.094)
NPLAZASGAR3	0.753**
	(0.129)
NPLAZASGAR4	0.919
	(0.170)
NPLAZASGAR5	0.888
	(0.135)
NPLAZASGAR6	1.458***
	(0.131)
NPLAZASGAR7	1.752***
	(0.190)
NPLAZASGAR8	1.395
	(0.347)
NPLAZASGAR9	3.800***
	(0.348)
NASEOS	1.156***
	(0.045)
SUCIO6	0.706***
	(0.078)
Observations	5,208
Note:	*p<0.1; **p<0.05; ***p<0.01

ANEXO 2: Resumen parámetros modelo final.

```
exp(coef(modelo5))
   CALEFAC2      CALEFAC3      CALEFAC4     TIPOAGUA2     TIPOAGUA3    AIREACOND6      INTERNET6      NDORMITO
  0.4633196     0.3001146     0.5827113     1.9570610     4.2814768     0.7702059     0.5692629     0.8051822
MALCOMUNIC6      COLEGIO6        BARES6         SUPER6     INGREHOG2     INGREHOG3     INGREHOG4     INGREHOG5
  1.2745391     0.7325693     1.4621696     0.5682916     0.9031273     1.2039077     1.3596686     1.6080124
  INGREHOG6     INGREHOG7     INGREHOG8     INGREHOG9   NPLANTASSOB   GASTUBERIA6      LUGTRAB1      LUGTRAB2
  2.0883361     2.1337519     4.8501483     1.5127242     1.2441109     0.5544567     0.8301451     1.0089132
   LUGTRAB3      LUGTRAB4      LUGTRAB5      LUGTRAB6 DELINCUENCIA6 LAVAVAJILLAS2     LAVADORA2     SECADORA2
  0.8376221     1.0975002     0.4548320     2.1001314     0.6747534     1.2161856     0.4503785     0.7302877
NPLAZASGAR1   NPLAZASGAR2   NPLAZASGAR3   NPLAZASGAR4   NPLAZASGAR5   NPLAZASGAR6   NPLAZASGAR7   NPLAZASGAR8
  0.7013557     0.7541756     0.7534244     0.9189829     0.8877810     1.4579729     1.7515038     1.3949183
NPLAZASGAR9       NASEOS        SUCIO6
  3.7997398     1.1560050     0.7062903
```

ANEXO 3: Resumen coeficientes de los parámetros.

```
Overall Statistics

Accuracy : 0.5622
95% CI : (0.5486, 0.5757)
No Information Rate : 0.4147
P-Value [Acc > NIR] : < 2.2e-16
```

ANEXO 4: Precisión del modelo.