

EL RETO DE LA DESPOBLACIÓN A TRAVÉS DE DATA SCIENCE

**ADRIANA BORRAJO VALIÑA
FEBRERO 2023**

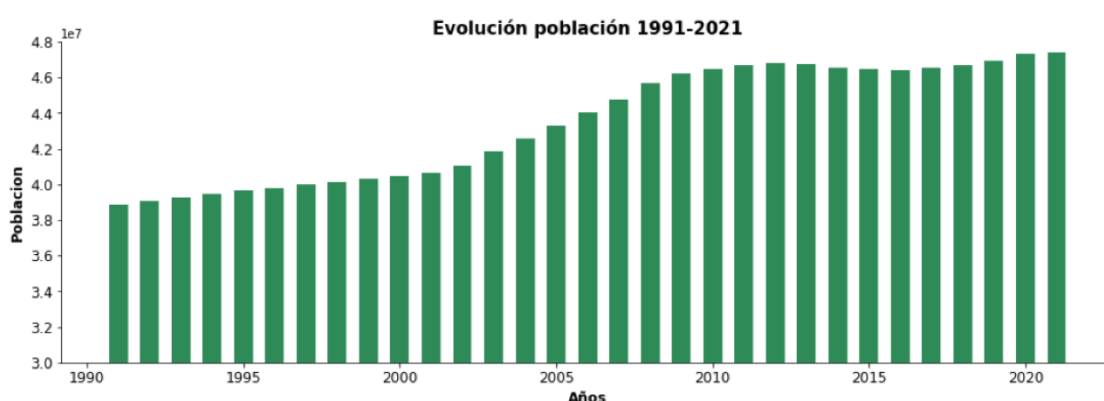
ÍNDICE

1.	INTRODUCCIÓN Y OBJETIVO	3
2.	REPOSITORIO DE LOS DATOS	5
3.	ESTUDIO DE LA DENSIDAD DE LA POBLACIÓN	11
4.	UNIÓN Y DEPURACIÓN DE LOS DATOS	13
4.1	UNIÓN DE LA INFORMACIÓN	13
4.2	LIMPIEZA DE LOS DATOS	13
4.2.1	FORMATO DE DATOS	13
4.2.2	VALORES DUPLICADOS	14
4.2.3	VALORES NULOS	14
5.	ANÁLISIS EXPLORATORIO DE LOS DATOS	16
5.1	ANÁLISIS PRELIMINAR	16
5.2	ANÁLISIS DE LA VARIABLE OBJETIVO	17
5.3	ANÁLISIS DESCRIPTIVO VARIABLES PREDICTORAS	18
5.4	ANÁLISIS UNI Y MULTIVARIANTE: REPRESENTACIÓN GRÁFICA	19
6.	CONSTRUCCIÓN DEL MODELO DE CLASIFICACIÓN	24
6.1	ESCALADO DE VARIABLES	24
6.2	DIVISIÓN CONJUNTO DE DATOS TRAIN Y TEST	24
6.3	MODELOS BASE Y EVALUACIÓN	25
6.4	OPTIMIZACIÓN DEL MODELO	27
6.5	EVALUACIÓN DEL MODELO OPTIMIZADO	28
6.6	IMPORTANCIA DE LAS CARACTERÍSTICAS	29
7.	TABLEAU	30

1. INTRODUCCIÓN Y OBJETIVO

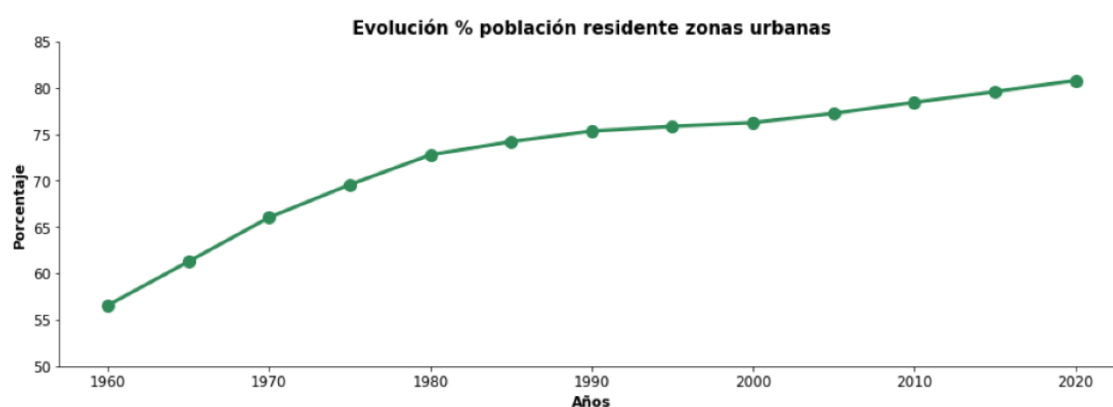
En un mundo donde cada vez vivimos más personas, la dinámica geográfica tiene una clara tendencia a la concentración. Habitamos en un número cada vez más reducido de núcleos urbanos como consecuencia de los beneficios que aporta la proximidad física: una mayor interacción social, servicios con menor coste por persona o un incremento de ganancias y productividad en las empresas.

España no es una excepción en el actual modelo territorial global. La población total ha experimentado un alto crecimiento en estos últimos 30 años:



Entre 1991 y 2021, el país pasa de 38,8 a 47,3 millones de habitantes. Unos 8,5 millones más de personas, lo que supone un crecimiento cercano al 22%.

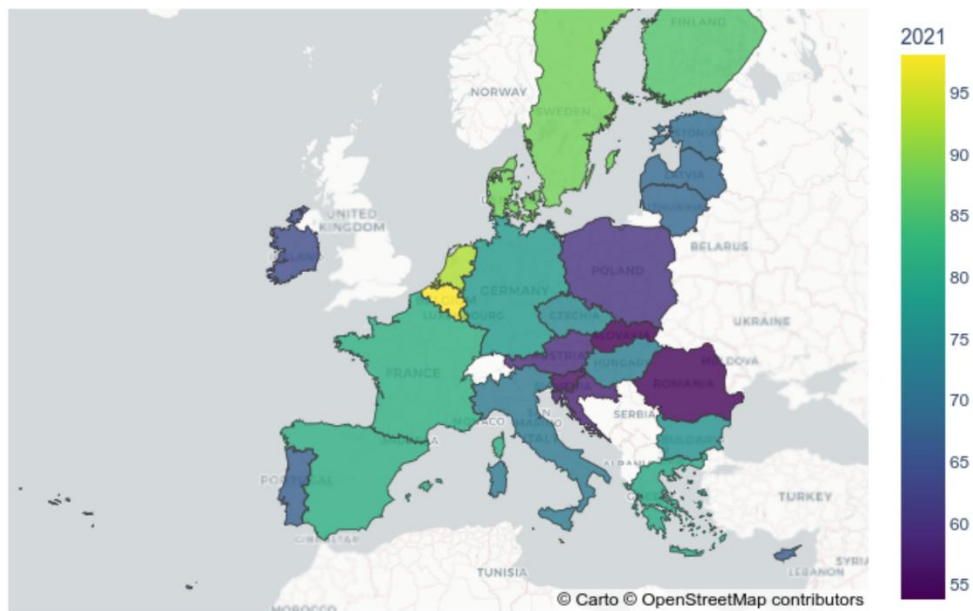
Sin embargo, estos datos esconden otra realidad, el crecimiento de la población va acompañado de otra tendencia al alza, la concentración de la población española en zonas urbanas:



Un claro proceso migratorio rural-urbano en el que la tasa de urbanización aumentó significativamente, desde los años 60, del 56% al 81% de 2021.

No obstante, este fenómeno no es exclusivo de España, estando bastante en línea con nuestro entorno, e incluso con unos datos ligeramente superiores a los del conjunto de la eurozona que se sitúa en el 74%:

% Población zonas urbanas UE



Afrontar el reto demográfico y reducir y eliminar las brechas territoriales no es posible sin el conocimiento e interpretación de la realidad de una región a través de su distribución, composición y estructura de su población.

Ante este problema, el objetivo de este proyecto es doble: efectuar un diagnóstico a 10 años con el que poder predecir la despoblación de un municipio en dicho periodo, e intentar buscar las variables más significativas para poder actuar sobre ellas.

Utilizaremos distintos modelos de clasificación supervisada para analizar e interpretar variables territoriales, demográficas, económicas y sociales con el fin de anticiparse a las necesidades futuras de cada municipio y prevenir posibles efectos negativos vinculados con la despoblación.

Como variable objetivo se utilizará la definición que fija la Comisión Europea, la cual establece un umbral crítico en los 12,5 hab/km². Todos los municipios que tengan una densidad por debajo de ese valor son catalogados como municipios despoblados.

2. REPOSITORIO DE LOS DATOS

En la carpeta *2.Data* se encuentran todos los datos recopilados para el proyecto. Dicha información está dividida en diferentes archivos estructurados en 35 bloques y organizados de la siguiente forma:

- **XX_Input_Nombre_Variable_Año:** Ficheros con los datos raíz extraídos directamente de la descarga de un Excel. Los bloques en los que se ha utilizado formato JSON no disponen de dicho fichero.
- **XX_Code_Nombre_Variable_Año:** .ipynb donde se tratan los datos para posteriormente guardarlos en un csv.
- **XX_Output_Nombre_Variable_Año:** Archivo de salida csv con los datos ya estructurados.

Para la búsqueda de la información se han consultado diversas fuentes. La más utilizada ha sido el INE (Instituto nacional de estadística) ya que recoge, produce y difunde el tipo de datos estadísticos que necesitamos para este estudio. Así mismo, también se han empleado otras instituciones públicas como, por ejemplo, el Ministerio de Transportes, de Asuntos Económicos o el de Sanidad.

Las dos primeras tablas de la carpeta *2.Data*, *01_Output_ProvMun_11* y *01_Output_ProvMun_21*, contienen los campos: *Nombre CCAA*, *Codigo CCAA*, *Nombre Provincia*, *Codigo Provincia*, *Nombre Municipio*, *Codigo Municipio*. Este último es el que se utilizará como primary key para juntar todas las tablas en un único DataFrame.

El denominado *Codigo Municipio* fue creado por el INE y su uso se extiende a través de la mayoría de las organizaciones estatales. Este código se compone de cinco dígitos: los dos primeros corresponden al código de la provincia y los tres restantes hacen referencia al municipio.

Aunque la exigencia de encontrar información que contenga la relación de los códigos de los municipios con las variables de estudio dificultó la búsqueda de los datos, fue un requisito indispensable para poder identificar de forma exclusiva a cada uno de los municipios que componen el territorio español. Se descartó desde un primer momento que el nexo de unión fuera el nombre de cada uno de los municipios al ser este un campo con mucha variabilidad en la escritura de los mismos.

También mencionar otro inconveniente: algunos valores del campo *Codigo Municipio* empiezan por cero. Para poder mantener dicha estructura al leer un csv, se ha impuesto a lo largo de todo el proyecto la condición de que este campo sea un string.

Las tablas creadas en esta fase previa a la fusión y posterior creación del DataFrame definitivo, tienen todas la siguiente estructura columnar:

Codigo Municipio	Variable 1	Variable 2	...

Los datos recopilados son todos pertenecientes al año 2011 a excepción de dos variables/tablas cuyos valores se han recogido tanto para el 2011 como para el 2021:

- **Códigos INE de las comunidades autónomas, provincias y municipios del territorio nacional:** *01_Output_ProvMun_11.csv* y *01_Output_ProvMun_21.csv*.

La estructura territorial de España ha sufrido bastantes variaciones a lo largo del tiempo. Por motivos que atienden a causas geográficas, sociales, económicas, culturales y de gestión, se han fusionado o dividido regiones, creado así nuevos municipios. En el año 2011 España estaba constituida por 8.116 municipios mientras que en el año 2021 este dato ha variado hasta la cifra actual de 8.131 municipios.

En consecuencia, además de la tabla *ProvMun_11*, también necesitamos *ProvMun_21*.

- **Población:** *02_Output_Poblacion_11.csv* y *02_Output_Poblacion_21.csv*

El motivo de necesitar la población de 2021 es que tal como explicaremos más adelante, será necesaria para calcular la variable objetivo de nuestro modelo de aprendizaje supervisado.

A continuación, se detalla la relación de tablas, variables y fuente de información que engloban toda la data:

Nombre tabla	Variables	Formato	Fuente
01_Input_ProvMun_11 01_Input_ProvMun_21 01_Code_ProvMun_11_21 01_Output_ProvMun_11 01_Output_ProvMun_21	Nombre CCAA Codigo CCAA Nombre Provincia Codigo Provincia Nombre Municipio Codigo Municipio	Excel	https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177031&menu=ultiDatos&idp=1254735976614
02_Code_Poblacion_11_21 02_Output_Poblacion_11 02_Output_Poblacion_21	Poblacion Total Poblacion Hombre Poblacion Mujer	Json	https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525
03_Input_Densidad_Poblacion_11_21 03_Code_Densidad_poblacion_11_21 03_Output_Densidad_Poblacion_11 03_Output_Densidad_Poblacion_21	Superficie Km2 Densidad Poblacion	Excel	http://centrodedescargas.cnig.es/CentroDescargas/index.jsp
04_Code_Altitud_Municipios 04_Output_Altitud_Municipios_11	Altitud Municipios	Excel	http://centrodedescargas.cnig.es/CentroDescargas/index.jsp
05_Code_Nacimientos_Matrimonios_Defunciones_11 05_Output_Nacimientos_Matrimonios_Defunciones_11	Nacimientos Matrimonios Defunciones	Json	https://www.ine.es/dynt3/inebase/es/index.htm?padre=3413
06_Input_Afiliados_SS_11 06_Code_Afiliados_SS_11 06_Output_Afiliados_SS_11	Total Regimen SS	Excel	https://www.seg-social.es/wps/portal/wss/internet/EstadisticasPresupuestosEstudios/Estadisticas/EST8/EST167/5b11b695-cf1c-4abe-8a12-ecd2d0d15271/2683/0985c546-dd0e-491b-b5e3-4e5e621c69b8
07_Input_Contratos_Laborales_11 07_Code_Contratos_Laborales_11 07_Output_Contratos_Laborales_11	Total Contratos Contratos Hombres Contrato Mujeres	Excel	https://datos.gob.es/es/catalogo/ea0021425-contratos-por-municipios
08_Input_Parados_11 08_Code_Parados_11 08_Output_Parados_11	Total Paro Registrado Parados Hombres Parados Mujeres	Excel	https://datos.gob.es/es/catalogo/ea0021425-paro-registrado-por-municipios

Nombre tabla	Variables	Formato	Fuente
09_Input_Numero_Establecimientos_11 09_Code_Numero_Establecimientos_11 09_Output_Numero_Establecimientos_11	Nº establecimientos Agricultura Nº establecimientos Industria Nº establecimientos Construcción Nº establecimientos Servicios	Excel	https://atlasau.mitma.gob.es/#c=indicador&i=estest.pie_est&s=2020&view=map4
10_Input_Numero_Centros_Sanitarios_11 10_Code_Numero_Centros_Sanitarios_11 10_Output_Numero_Centros_Sanitarios_11	Nº centros sanitarios	Excel	https://www.sanidad.gob.es/estadisticas/microdatos.do
11_Input_Numero_Centros_Educativos_11 11_Code_Numero_Centros_Educativos_11 11_Output_Numero_Centros_Educativos_11	Nº Centros Educativos	Excel	https://www.educacionyfp.gob.es/servicios-al-ciudadano/estadisticas/no-universitaria/centros.html
12_Input_Transacciones_Inmobiliarias_11 12_Code_Transacciones_Inmobiliarias_11 12_Output_Transacciones_Inmobiliarias_11	Número de transacciones anuales 2011	Excel	https://atlasau.mitma.gob.es/#c=indicador&i=vivtrana.trans001&s=2020&view=map4
13_Code_Establecimientos_Colectivos_11 13_Output_Establecimientos_Colectivos_11	Nº Residencias Mayores Nº Instituciones Religiosas y Militares	JSON	https://www.ine.es/jaxi/Tabla.htm?path=/t20/e244/colectivos/p05/I0/&file=1mun00.px&L=0
14_Code_Nacidos_Extranjero_11 14_Output_Nacidos_Extranjero_11	Nº Hombres Extranjeros Nº Mujeres Extranjeros Nº Total Extranjeros	JSON	https://www.ine.es/jaxi/Tabla.htm?path=/t20/e244/avance/p02/I0/&file=2mun00.px&L=0
15_Input_Nacidos_otra_CCAA_11 15_Output_Nacidos_otra_CCAA_11	Nº Hombres Otra CCAA Nº Mujeres Otra CCAA Nº Total Otra CCAA	JSON	https://www.ine.es/jaxi/Tabla.htm?path=/t20/e244/avance/p02/I0/&file=2mun00.px&L=0
16_Code_Edificios_Inmuebles 16_Output_Edificios_Inmuebles	Nº edificios Nº inmuebles	JSON	https://www.ine.es/jaxi/Tabla.htm?path=/t20/e244/edificios/p04/I0/&file=1mun00.px&L=0
17_Code_Viviendas_11 17_Output_Viviendas_11	Nº Viviendas	JSON	https://www.ine.es/jaxi/Tabla.htm?path=/t20/e244/viviendas/p06/I0/&file=9mun00.px&L=0
18_Input_Edad_11 18_Code_Edad_11 18_Output_Edad_11	Edad < 16 16 ≤ Edad < 64 Edad ≥ 65	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2

Nombre tabla	Variables	Formato	Fuente
19_Input_Estado_Civil_11 19_Coode_Estado_Civil_11 19_Output_Estado_Civil_11	Soltero/a Casado/a Separado/a Divorciado/a Viudo/a	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
20_Input_Nivel_Estudios_11 20_Code_Nivel_Estudios_11 20_Output_Nivel_Estudios_11	Sin Estudios Con Estudios	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
21_Input_Jubiados_11 21_Code_Jubiados_11 21_Output_Jubiados_11	Nº de Jubilados	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
22_Input_Empresarios_11 22_Code_Empresarios_11 22_Output_Empresarios_11	Empresario que emplea personal Empresario que no emplea personal	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
23_Input_Numero_Hijos_11 23_Code_Numero_Hijos_11 23_Output_Numero_Hijos_11	1 hijo/a 2 hijos/as 3 hijos/as 4 o más hijos/as	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
24_Input_Alquileres_11 24_Code_Alquileres_11 24_Output_Alquileres_11	Nº de Alquileres	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
25_Input_Tipos_Hogar_11 25_Code_Tipos_Hogar_11 25_Output_Tipos_Hogar_11	Hogares Unipersonales Hogares Familiares	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
26_Input_Personas_Hogar_11 26_Code_Personas_Hogar_11 26_Output_Personas_Hogar_11	1 persona 2 personas 3 personas 4 personas 5 personas 6 o más personas	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2

Nombre tabla	Variables	Formato	Fuente
27_Input_Edad_Edificios_11 27_Code_Edad_Edificios_11 27_Output_Edad_Edificios_11	Edad Media Edificios	Excel	https://atlasau.mitma.gob.es/#c=indicador&i=tipant.tant005&s=2020&view=map4
28_Input_Uso_Suelo_11 28_Code_Uso_Suelo_11 28_Output_Uso_Suelo_11	Uso Industrial Uso Residencial Uso Terciario Otros Usos	Excel	https://atlasau.mitma.gob.es/#c=indicador&i=tipant.tant005&s=2020&view=map4
29_Input_Acceso_Internet_11 29_Code_Acceso_Internet_11 29_Output_Acceso_Internet_11	Nº Hogares si Internet Nº Hogares no Internet	Excel	https://www.ine.es/censos2011/tablas/Wizard.do?WIZARD=1&reqCode=paso2
30_Input_Parque_Vehiculos_11 30_Code_Parque_Vehiculos_11 30_Output_Parque_Vehiculos_11	Nº Vehiculos	Excel	https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces
31_Input_Censo_Conductores_11 31_Code_Censo_Conductores_11 31_Output_Censo_Conductores_11	Nº Conductores	Excel	https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces
32_Input_Deuda_Ayuntamientos_11 32_Code_Deuda_Ayuntamientos_11 32_Output_Deuda_Ayuntamientos_11	Deuda (Miles de €)	Excel	https://serviciotelematicosext.hacienda.gob.es/SGFAL/CONPREL
33_Input_Impuestos_11 33_Code_Impuestos_11 33_Output_Impuestos_11	Total impuestos (Miles de €)	Excel	https://serviciotelematicosext.hacienda.gob.es/SGFAL/CONPREL
34_Input_Explotaciones_Agricolas_11 34_Code_Explotaciones_Agricolas_11 34_Output_Explotaciones_Agricolas_11	Superficie (Ha) Agricolas Nº Explotaciones Agricolas	Excel	https://www.ine.es/CA/Inicio.do
35_Input_Explotaciones_Ganaderas_11 35_Code_Explotaciones_Ganaderas_11 35_Output_Explotaciones_Ganaderas_11	Nº Explotaciones Ganaderas	Excel	https://www.ine.es/CA/Inicio.do

3. ESTUDIO DE LA DENSIDAD DE LA POBLACIÓN

La comisión europea utiliza la variable densidad para clasificar el declive de las regiones que están escasamente pobladas respecto de las que no.

Según su criterio, un territorio despoblado es aquel que tiene una densidad inferior a 12,5 habitantes por km², mientras que un territorio poblado es aquel que tiene una densidad igual o superior a los 12,5 habitantes por Km².

Siguiendo este enfoque, a través de un análisis geoespacial, realizamos un primer diagnóstico de cómo se distribuye la densidad de la población en los municipios de España entre los años 2011 y 2021.

Para ello, se siguen los siguientes pasos:

Notebook *Umbral_Densidad_Poblacion.ipynb*:

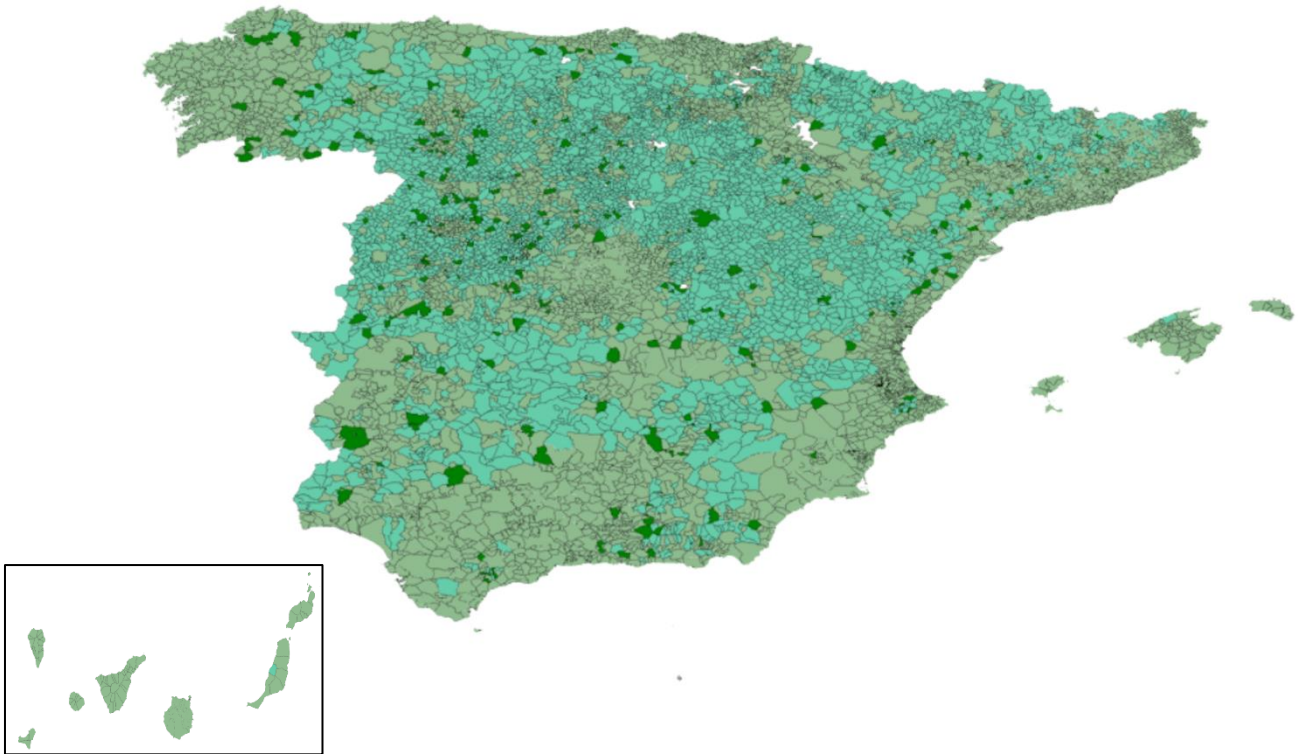
- Se utilizan las tablas, *03_Output_Densidad_Poblacion_11.csv* y *03_Output_Densidad_Poblacion_21.csv*, donde previamente se ha calculado a través de la superficie de cada municipio y el número de habitantes, la densidad en habitantes/km² para los años 2011 y 2021.
- Se clasifican tanto a los municipios de 2011 como de 2021 en poblados o despoblados siguiendo el criterio de la Comisión Europea:
 - 2011:
 - Nº de municipios poblados: 4.411
 - Nº de municipios despoblados: 3.705
 - 2021:
 - Nº de municipios poblados: 4.184
 - Nº de municipios despoblados: 3.947
- Se realiza un merge de las dos tablas y se crean dos nuevas columnas:
 - **Densidad:** Si en 2011 era ya un municipio despoblado, toma la densidad de 2011. En cambio, si en 2011 estaba poblado, toma el valor de la densidad de 2021.
 - **Umbral:** Si el municipio ya estaba despoblado desde 2011, esta variable toma el valor 'Despoblado11', si en 2011 estaba poblado pero en 2021 ya no, toma el valor 'Despoblado21' y si en 2021 está poblado, toma el valor 'Poblado'.
- Dividimos el DataFrame en dos: Uno con los municipios de Canarias y otro con el resto de los municipios de España. Los guardamos en dos csv: *Canarias* y *Peninsula*.

Notebooks *Mapa_Densidad_Canarias.ipynb* y *Mapa_Densidad_Peninsula.ipynb*:

- Se descargan del Instituto Geográfico Nacional dos ficheros Shapefile:
 - Uno con las entidades geométricas de la península, Baleares, Ceuta y Melilla.
 - El otro con las entidades geométricas de Canarias.
- Junto con la columna *Umbral* de los ficheros *Canarias.csv* y *Peninsula.csv*, y utilizando el campo *Codigo Municipio* como punto de unión, se crean dos mapas (península y Canarias),

que representan a los municipios que están despoblados desde 2011, a los que se convierten en despoblados en 2021 y a los que en 2021 siguen siendo poblados.

Se obtiene el siguiente mapa:



Los territorios verde botella representan a los municipios poblados, con una densidad igual o superior a 12,5 habitantes/km². Las zonas con una tonalidad verde agua, son los municipios que están despoblados desde 2011, y las áreas en color verde más intenso, son municipios que desde 2011 hasta 2021 han sufrido una transformación de áreas pobladas a despobladas, es decir, han pasado de tener una densidad igual o superior a 12,5 habitantes/km², a tener una densidad inferior a dicha cifra.

A través de este análisis exploratorio se puede observar que España tiene una distribución territorial con grandes diferencias. La población presenta un elevado grado de concentración tanto en el litoral como en las zonas urbanas, mientras que las áreas más despobladas corresponde a zonas del interior donde la mayoría se distribuye entre las dos Castillas, Aragón y parte de Andalucía. No obstante, todas las comunidades autónomas disponen de territorios señalados, en mayor o menor medida.

Llama la atención contra todo pronóstico, que áreas con mucho rural, como puede ser Galicia o Asturias, apenas tienen municipios despoblados.

4. UNIÓN Y DEPURACIÓN DE LOS DATOS

En esta fase se procede a la integración de toda la información recopilada previamente de múltiples fuentes de datos, así como a la limpieza de los mismos.

4.1 UNIÓN DE LA INFORMACIÓN

Es el momento de reunir todos los datos y crear un único DataFrame con el que vamos a trabajar en las posteriores etapas del proyecto.

Para ello, a través de un bucle, se leen todos los *Output.csv* que están en la carpeta *2.Data* mientras se guardan en una lista.

Apuntar que, previamente, en la fase de creación de los *Output.csv*, ya se trabajó en una limpieza de columnas, eliminando todas la que no se necesitan, se renombraron variables para que estas sean más legibles y se añadió a todas las tablas una primera columna con el *codigo municipio*, con el fin de organizar y estructurar de la mejor forma los datos y facilitar la integración de toda la información en una única tabla.

Utilizando las funciones *reduce* y *lambda* y la característica de que todas las tablas constan del *Codigo Municipio*, se realiza un *merge* a lo largo de la lista de *.csv* que se acaba de crear.

Se obtiene así el DataFrame *Despoblacion*, que consta de unas seis primeras columnas con información geográfica; *Nombre CCAA*, *Codigo CCAA*, *Nombre Provincia*, *Codigo Provincia*, *Nombre Municipio*, *Codigo Municipio*, y el resto de columnas que corresponden a variables demográficas, económicas, sociales...

4.2 LIMPIEZA DE LOS DATOS

El objetivo de esta parte es corregir y eliminar las posibles anomalías en la estructura de los datos, descartando los valores falsos y las inconsistencias que se arrastran de las fuentes de datos.

4.2.1 FORMATO DE DATOS

Realizamos el cambio de todas las variables a formato numérico a excepción de las 6 primeras columnas, correspondientes a las variables de ámbito geográfico, que seguirán teniendo formato *object*.

Antes de poder transformar todas las variables en numéricas, se detecta que en unas 30 columnas, correspondientes a información extraída del INE, toman el valor *asterisco: **.

Según la legislación sobre secreto estadístico, toda organización está obligada a difundir la información estadística de manera que no pueda saberse, ni siquiera indirectamente, a quién corresponde un dato particular.

Siguiendo la normativa vigente, el INE substituye los valores muestrales menores que 5 por un asterisco. Para solucionar este inconveniente, se ha decidido substituir los asteriscos por un 2, valor medio de los datos muestrales faltantes.

Con esta limpieza, ya se realiza el cambio, de la sexta fila en adelante, a formato numérico: float64.

4.2.2 VALORES DUPLICADOS

El total de municipios en España en 2011 era de 8.116, coincidiendo con las dimensiones del DataFrame.

No obstante, se comprueba que efectivamente no hay duplicados y que cada una de las filas de la tabla corresponden a cada uno de los municipios que conforman el territorio español.

4.2.3 VALORES NULOS

En un primer paso se detectan las columnas que contienen valores nulos y cuántos hay en cada una de ellas. Posteriormente se analizan dichas variables y se intenta descubrir sus causas; falta de información, no haber podido estimar la variable, un error de entrada, o bien un fenómeno externo que impide la recolección de datos.

En función de la casuística en la que nos encontremos, se podrá deducir el valor y completarlo, o se tendrá que eliminar el registro por ausencia o desconocimiento de dicho valor.

Se detallan a continuación las variables con valores nulos encontradas en el DataFrame *Despoblacion* y el criterio que se ha seguido:

- **Total Régimen SS:** Hay un total de 36 valores nulos. El motivo es que la tabla correspondiente a la afiliación a la Seguridad Social no contiene datos de estos 36 municipios. Decidimos eliminar las 36 filas correspondientes.
- **Nº de transacciones anuales 2011:** Hay un total de 4 nulos. En 2016 se fusionan cuatro municipios; Cercedo-Cotobade, Oza-Cesuras. Casualmente, en la tabla transacciones, los municipios aparecen fusionados. El valor de Cercedo-Cotobade es nulo, y por tanto, se eliminan ambos municipios del DataFrame *Despoblacion*. En cambio, el número de transacciones de Oza-Cesuras es de 27. Se divide este dato según la proporcionalidad de población de cada municipio y se traslada al DataFrame.
- **Nº Establecimientos Agricultura, Industria, Construcción, Servicios, Centros Sanitarios, Centros Educativos, Residencias Mayores, Instituciones Religiosas y Militares:** Los valores nulos corresponden a la falta de dichos establecimientos en los municipios correspondientes. Substituimos los valores nulos por cero.
- **Nº extranjeros por municipio:** Hay 722 nulos en hombres, 730 en mujeres y 466 en el total de extranjeros. Si nos fijamos en los 466 nulos, vemos que son municipios con una población muy pequeña. Deducimos que la falta de valor significa la ausencia de extranjeros en estos municipios. Substituimos los valores nulos por cero.
- **Nº de personas otras comunidades autónomas:** Lo mismo que el caso anterior, hay 211 nulos en hombres, 280 en mujeres y 96 en el total. Substituimos los valores nulos por cero.
- **Nº de vehículos:** 8 de los 9 valores nulos coinciden con que también son nulos los valores de la variable número de conductores. Asignamos el valor cero en ambos campos. El municipio

faltante tiene 2 conductores. Como la población es muy pequeña, 230 habitantes, determinamos que el número de vehículos es cero.

- **Nº de conductores:** Hay 3 municipios con valores nulos y con número de vehículos: 579, 2 y 6. De estos dos últimos asumimos que el número de conductores es cero. Para el municipio con 579 vehículos, se ha encontrado el dato del número de conductores en 2012: 351. Sustituimos el nulo por dicho valor.
- **Total impuestos (Miles €):** Hay 5 municipios con valores nulos. Se han buscado estos datos en el Ministerio de Hacienda y aparecen con valor cero. Sustituimos los nulos por cero.
- **Superficie (Ha) Agrícolas, Nº Explotaciones Agrícolas, Nº Explotaciones Ganadera:** Los nulos son municipios donde no hay actividad. Sustituimos los nulos por cero.
- **Edad Media Edificios y uso del suelo:** hay un total de 522 y 517 nulos respectivamente. En un primer momento se valoró eliminar estos registros. El problema es que, de ellos, 248 son municipios del País Vasco y 271 de Navarra. Estas dos CCAA tienen 251 y 271 municipios respectivamente. Si se eliminan los nulos, nos quedamos sin información de estas dos áreas. Se decide excluir del DataFrame las variables edad media y uso del suelo.

Finalmente, con todas estas modificaciones, el tamaño del DataFrame *Despoblacion* es de 8.077 filas por 75 columnas.

Guardamos el DataFrame depurado en formato csv: *Despoblacion_11*.

5. ANÁLISIS EXPLORATORIO DE LOS DATOS

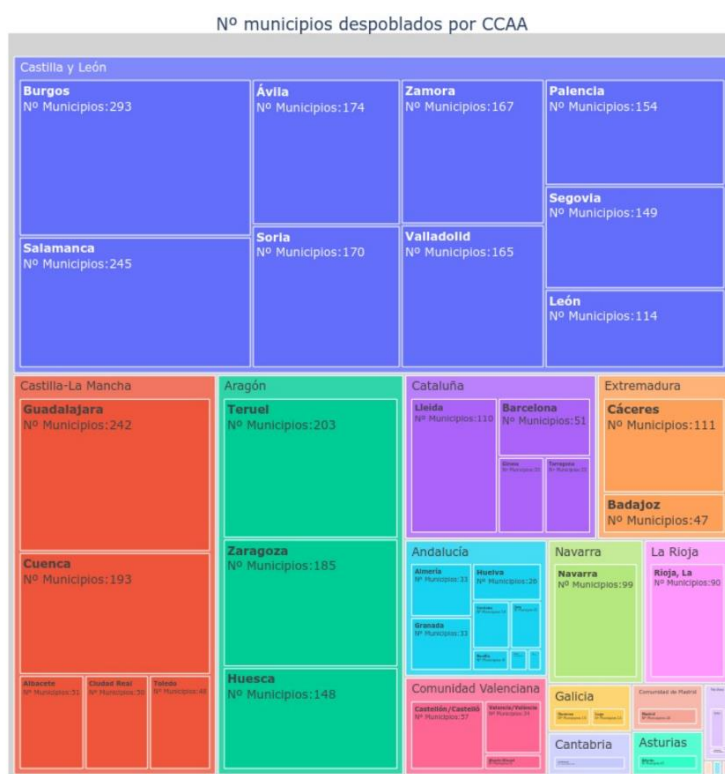
El comportamiento de los datos es un aspecto clave del análisis y de la construcción de modelos. Es importante realizar una primera aproximación del significado y la relevancia de cada variable respecto al problema planteado.

El proyecto se basa en la predicción a 10 años de la despoblación de un municipio, siendo las variables predictoras datos demográficos relativos a 2011, y la variable objetivo una variable categórica que clasifica a los municipios según el valor de la densidad que tienen en el año 2021: 1 = Despoblados ($< 12,5$ habitantes/km²), 0 = Poblados ($\geq 12,5$ habitantes/km²).

En el estudio que se expone a continuación, se realiza una exploración preliminar y un análisis descriptivo.

5.1 ANÁLISIS PRELIMINAR

Un primer análisis se centra en conocer que CCAA y provincias concentran más municipios despoblados en 2011:



Observamos que Castilla y León es la CCAA más afectada por la despoblación y el declive demográfico. En concreto, la provincia de Burgos lidera el ranking con 293 municipios despoblados de un total de 371. En un segundo puesto está Salamanca con 245, y muy cerquita, en el tercer puesto, tenemos a Guadalajara cuya CCAA, Castilla-La Mancha, es la siguiente en abarcar más municipios despoblados.

Continuando con otros datos demográficos:

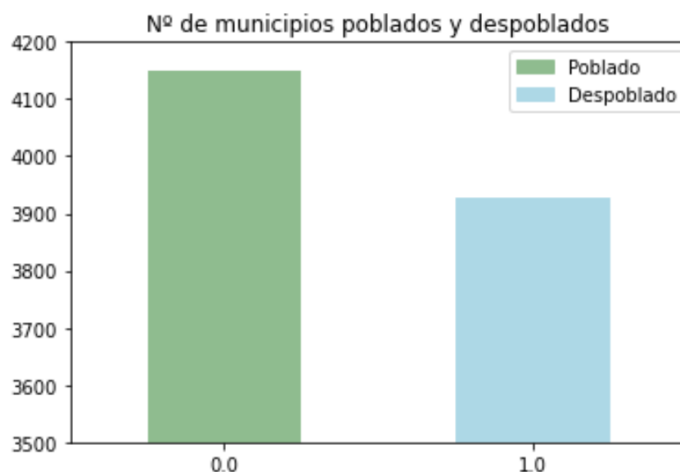
- **Población total, % de hombres, % de mujeres y densidad media en los años 2011 y 2021:**
 - Mínimo crecimiento demográfico, de 47,11 millones de habitantes en 2011 a 47,38 en 2021.
 - Leve incremento en el porcentaje de mujeres, de 50,7% en 2011 a 51% en 2021.
 - Aumento de la densidad, de 174,28 en 2011 a 178,60 en 2021.
- **Grado de concentración rural 2011 y 2021:**
 - En 2011, el 45,65 % de la superficie española era ya territorio despoblado y vivían en estas zonas tan solo el 2,74 % de la población.
 - En 2021 la situación se agrava, siendo el 48,65% de la superficie. Se produce así una disminución en el grado de ruralidad de los municipios y un incremento en el grado de aglomeraciones urbanas.
- **Municipios despoblados 2021 y su situación en 2011:**
 - 3.926 municipios despoblados en 2021 de los cuales 261 no estaban despoblados en 2011.
 - Malpartida de Plasencia, municipio de Cáceres, con mayor descenso de densidad: - 22 hab/km².
 - Calañas, municipio de Huelva, con un mayor descenso de población: -1.424 habitantes.
- **Municipios poblados 2021 y su situación en 2011:**
 - 4.150 municipios poblados en 2021 de los cuales, 21 estaban despoblados en 2011.
 - Buenavista, municipio de Salamanca, con mayor ascenso de densidad: 4,4 hab/km².
 - Golmayo, municipio de Soria, con un mayor ascenso de población: 711 habitantes.
- **Municipios con mayor población:**
 - 6 municipios por encima del medio millón de habitantes en 2011.
 - Madrid con mayor incremento de población: 40.370 habitantes.
 - Barcelona con mayor incremento de densidad: 211,24 hab/km².
 - Descenso de población y densidad en Valencia y Sevilla.

5.2 ANÁLISIS DE LA VARIABLE OBJETIVO

La variable objetivo llamada *Umbral*, toma el valor 0 para municipios poblados y 1 para municipios despoblados.

Contabilizamos cuantos registros hay de uno y de otro:

- 0/Despoblados: 4.150
- 1/Poblados: 3.926



5.3 ANÁLISIS DESCRIPTIVO VARIABLES PREDICTORAS

En una primera visualización, se observa un mismo patrón común en todas las variables: Gran distancia entre la media y el valor máximo:

	Poblacion Total	Poblacion Hombre	Poblacion Mujer	Superficie Km2	Densidad Poblacion	Altitud Municipios	Nacimientos	Matrimonios	Defunciones	Total Regimen SS	Total Contratos	Contratos Hombres	Contratos Mujeres	Total Paro Registrado
count	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00	8076.00
mean	5834.51	2878.63	2955.88	62.29	174.28	613.02	58.18	19.99	47.72	2087.54	1784.17	974.55	809.62	546.76
std	47544.00	22533.50	25018.15	92.26	879.56	343.71	475.23	179.87	397.31	24489.01	17244.55	8670.59	8613.03	3783.14
min	5.00	3.00	2.00	0.03	0.29	1.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00
25%	179.00	96.00	82.00	18.48	5.55	334.00	0.00	0.00	2.00	34.00	19.00	12.00	5.00	8.00
50%	582.00	306.00	280.00	34.94	15.04	666.00	3.00	1.00	7.00	132.00	100.00	59.00	36.00	38.00
75%	2492.25	1265.25	1229.00	69.23	57.04	857.00	21.00	8.00	26.00	676.00	612.75	359.00	238.25	199.25
max	3265038.00	1529842.00	1735196.00	1750.23	21505.91	1695.00	32938.00	12941.00	26544.00	1736667.00	1103251.00	557850.00	545401.00	222103.00

El motivo son los grandes núcleos urbanos, como por ejemplo Madrid, Barcelona, Málaga... donde los valores de todas las variables se disparan frente a la mayoría de los municipios. Observando adicionalmente las desviaciones y los valores mínimos, vemos que hay un rango muy amplio en el conjunto de los datos.

Si nos centramos en la densidad de población, vemos que también tiene una importante variación, con un mínimo de 0,29 hab/Km² a un máximo de 21.505,91 hab/Km². La media es de 174,28 hab/Km² con un nivel de desviación alto de 879,56.

A continuación, las 5 variables con mayor rango en los datos:

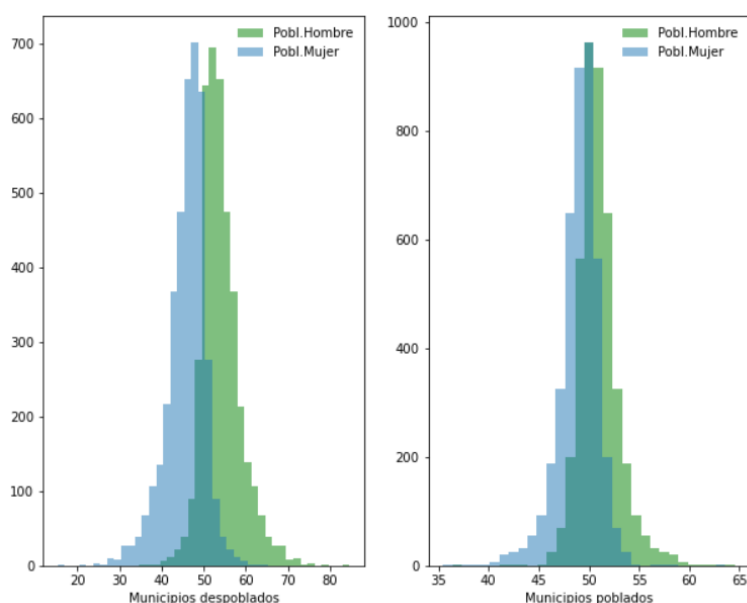
	Deuda (Miles de €)	Poblacion Total	Hogares Familiares	Con Estudios	16 =< Edad =<64
count	8076.00	8076.00	8076.00	8076.00	8076.00
mean	3484.98	5834.51	5222.77	3565.47	3851.34
std	74299.00	47544.00	41404.12	31908.63	31085.92
min	0.00	5.00	2.00	4.00	2.00
25%	0.00	179.00	145.00	82.00	100.00
50%	62.00	582.00	492.50	300.00	350.00
75%	586.25	2492.25	2210.00	1365.00	1580.00
max	6347900.00	3265038.00	2806765.00	2208075.00	2112935.00

5.4 ANÁLISIS UNI Y MULTIVARIANTE: REPRESENTACIÓN GRÁFICA

En este apartado, a través de la representación gráfica, se observará la relación entre sus características y la variable objetivo: Poblado o Despoblado.

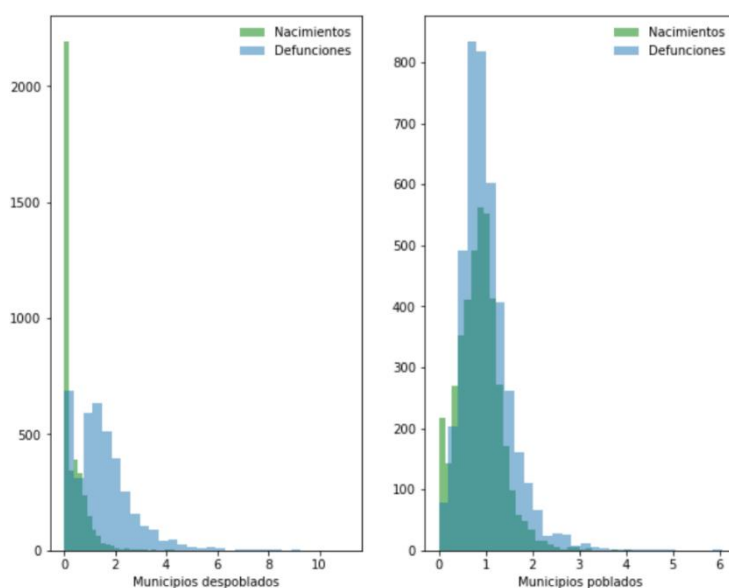
Uno de los problemas detectados en el apartado anterior es la gran extensión de nuestro rango de datos. Como ayuda para una correcta visualización, vamos a utilizar, para variables relacionadas con la población, el % de la variable respecto a la población total de cada municipio, y para el resto de variables, la escala logarítmica. Estas transformaciones nos van a permitir mostrar una amplia gama de puntos de datos sin que los valores muy pequeños y/o grandes se compriman en los dos extremos del gráfico.

- **Población hombre y mujer segmentada por municipios despoblados y poblados:**



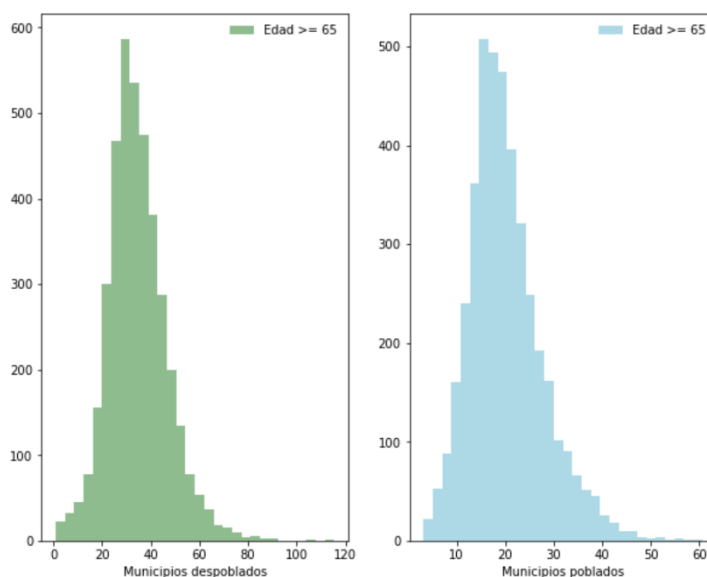
Observamos que, en los municipios poblados, el pico de frecuencia de hombres y mujeres prácticamente coincide en el 50%. En municipios despoblados, en hombres, el pico se alcanza en torno al 55%, mientras que en mujeres está alrededor del 48%. Además, la distribución en los municipios despoblados es más dispersa, teniendo más valores extremos, el izquierdo correspondiente a la población mujeres y el derecho correspondiente a hombres. En municipios poblados, aunque hay pequeñas colas a ambos lados, las distribuciones están bastante centradas.

- **Nacimientos y defunciones segmentado por municipios despoblados y poblados:**



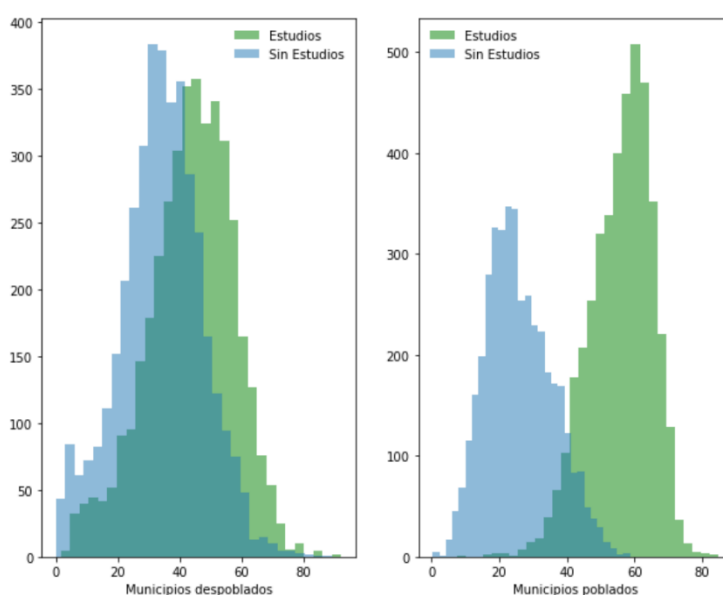
La mayor frecuencia de Nacimientos en municipios despoblados se concentra en el 0 mientras que las defunciones están en torno al 0.5 porcentual. La cola de la distribución de Nacimientos se queda en el 4 mientras que en Defunciones casi llega al 9. Respecto al grupo de municipios poblados, la variable Defunciones tiene un sesgo más pronunciado hacia la izquierda donde su máximo no llega al 1% mientras que la relativa a Nacimientos es mucho más dispersa y su pico sí que llega al 1.

- **Población mayor de 65 años segmentada por municipios despoblados y poblados:**



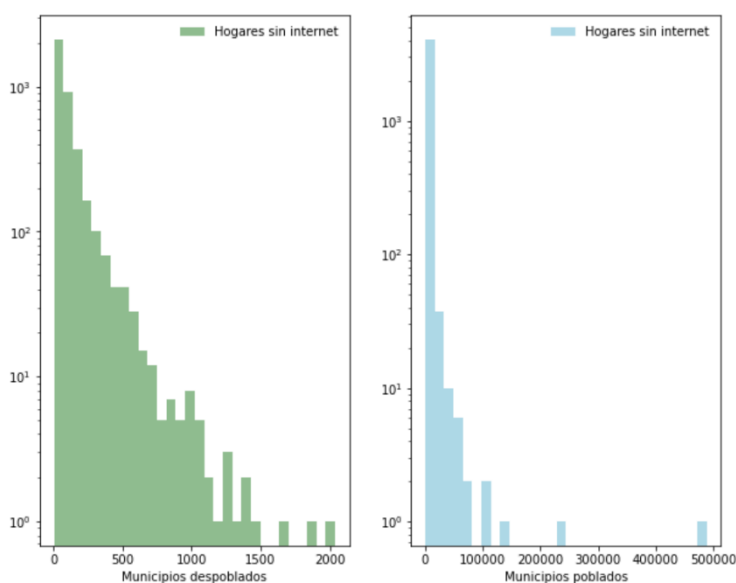
Las frecuencia más alta en municipios despoblados está prácticamente en el 30% de la población mientras que en municipios poblados no llega al 20%. La cola en municipios despoblados es mucho más dispersa.

- **Población con y sin estudios segmentada por municipios despoblados y poblados:**



En municipios poblados, la máxima frecuencia se encuentra por encima del 60% de la población y su dispersión por la derecha supera el 80%. En la población sin estudios, el pico de frecuencia baja hasta el 20 y la cola hacia la derecha, mucho más progresiva, no llega al 60%. En los municipios poco poblados, la distribución de la población con estudios es muy dispersa y con frecuencia máxima más baja que en poblados, en torno a un 50% de la población. Respecto a las personas sin estudios, llama la atención los valores por encima del 85%. Corresponde a Torrecuadrilla, municipio de Guadalajara y La Bastida, municipio de Salamanca.

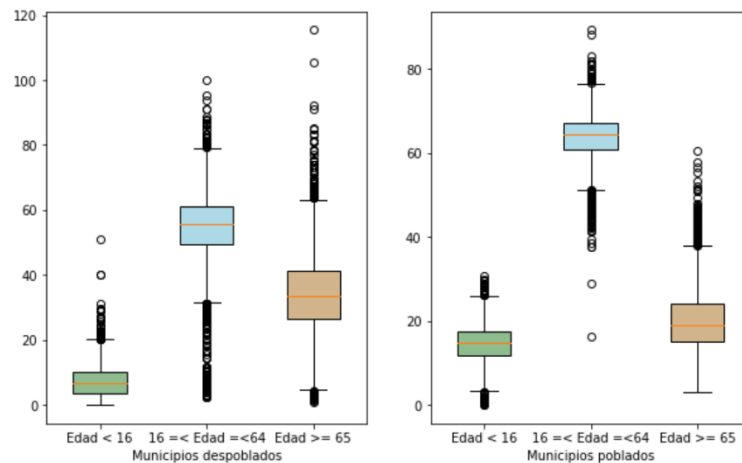
- **Nº de hogares sin internet por municipios despoblados y poblados:**



Como era de esperar, en municipios despoblados la concentración del número de hogares sin internet es más repartida, estando el grueso en torno a las 100 viviendas. En los

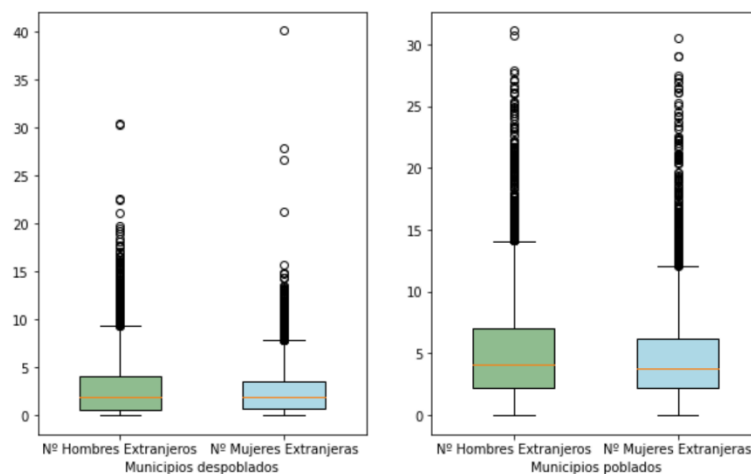
municipios poblados, llama la atención el pico cercano a 500.000. Corresponde a Madrid Capital donde hay unas 488.000 viviendas sin internet.

- **Grupos de edades segmentadas por municipios despoblados y poblados:**



Tanto en poblados como despoblados, las variables se comportan igual, los casos predominantes son las personas entre 16 y 64 años. En cambio, si realizamos la comparativa de estas tres variables dentro de los grupos de municipios despoblados y poblados, ya no hay tantas similitudes. Mientras que en los despoblados la mediana de menos de 16 años está en torno al 9%, la mediana de mayores de 65 está en un 30%. En cambio, en municipios poblados, la diferencia entre estas dos variables es más pareja, mediana de un 15% en menores de 16 y de un 20% en mayores de 65 años. En ambos casos, despoblados y poblados, hay muchos valores extremos.

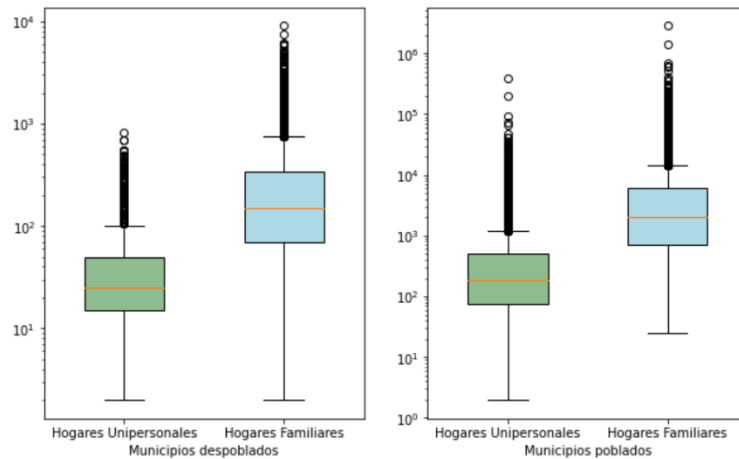
- **Número de personas extranjeras segmentadas por municipios despoblados y poblados:**



Mucha más cantidad de variabilidad en zonas pobladas que despobladas. Más hombres que mujeres en municipios poblados siendo la mediana más o menos pareja llegando casi al 5%.

En municipios despoblados, llama la atención el valor atípico en mujeres extranjeras. Corresponde a Fuente el Olmo de Fuentidueña, municipio de Segovia, con 69 mujeres extranjeras con población total femenina de 110.

- **Número de hogares unipersonales y familiares segmentados por municipios despoblados y poblados:**



Como también es de esperar, hay más hogares en municipios poblados que despoblados. En ambos casos, bigotes inferiores muy alargados, los valores mínimos están bastante alejados del primer cuartil. Un mayor recorrido en el rango intercuartílico en los hogares familiares de los municipios despoblados

6. CONSTRUCCIÓN DEL MODELO DE CLASIFICACIÓN

Utilizaremos el aprendizaje automático supervisado para la creación de un modelo de clasificación binaria cuyo objetivo es predecir, en un periodo a 10 años, a qué categoría, poblado o despoblado, pertenece una nueva observación.

Para ello, nuestro primer paso será construir y entrenar 5 modelos diferentes utilizando los siguientes algoritmos disponibles todos ellos en la librería Scikit-Learn:

- Regresión logística
- Support Vector Machines (SVM)
- Random Forest
- Naive Bayes
- K-vecinos más cercanos (KNN)

Posteriormente vamos a evaluar y comparar cual es el mejor modelo para nuestro caso concreto de estudio. Por último, optimizaremos el modelo ajustando los hiper parámetros con GridSearch.

Como se ha comentado anteriormente, la variable objetivo *Umbral* toma el valor 1 cuando la densidad de población en 2021 está por debajo de 12,5 hab/km² (municipio despoblado) y 0 cuando es igual o está por encima de 12,5 hab/km² (municipio poblado). Por tanto, 0 y 1 son las etiquetas que predecirá nuestro modelo. La mayoría de las otras columnas (*Población Total*, *Población hombre*, *Población Mujer*, etc.) son las características que usaremos para predecir la etiqueta *Umbral*.

6.1 ESCALADO DE VARIABLES

Para evitar problemas de estabilidad y mejorar el rendimiento del modelo, vamos a escalar las variables numéricas. De esta manera impedimos que algunas variables que toman valores elevados no sean utilizadas por el algoritmo como predictoras principales. En nuestro caso en concreto utilizaremos el método MinMaxScaler del paquete sklearn.preprocessing.

Descartamos en el escalado: *Nombre CCAA*, *Codigo CCAA*, *Nombre Provincia*, *Codigo Provincia*, *Nombre Municipio*, *Codigo Municipio*, *Umbral*.

6.2 DIVISIÓN CONJUNTO DE DATOS TRAIN Y TEST

Utilizaremos la función sklearn.model_selection.train_test_split para dividir nuestro DataFrame en dos bloques: datos de entrenamiento y datos de validación del modelo. Usaremos solo una parte para entrenar, y retener el resto para probar el modelo entrenado. Eso nos permitirá comparar las etiquetas pronosticadas con las ya conocidas.

Pasamos como parámetros de la función:

- Las variables independientes eliminando *Nombre CCAA*, *Codigo CCAA*, *Nombre Provincia*, *Codigo Provincia*, *Nombre Municipio*, *Codigo Municipio*. También descartamos *Superficie Km2* y *Densidad Población* por ser características que están relacionadas directamente con la variable objetivo, pudiendo tener una influencia demasiado fuerte en el modelo.
- La variable dependiente *Umbral*

- **Train_size:** con un tamaño del conjunto de entrenamiento del 0,8, y por extensión, de un 0,2 para test. En nuestro caso particular, de las 8.076 filas, 6.460 van a estar en el subconjunto train y 1.616 pertenecerán al subconjunto test.
- **Random_state:** con un valor de 42 garantizando que la semilla del generador aleatorio siempre sea la misma.
- **Stratify:** indicando como dato la columna *Umbral* para asegurar que la proporción de las dos clases del dataset original estén representadas por igual tanto en train como en test.

6.3 MODELOS BASE Y EVALUACIÓN

En este paso entrenamos un modelo de referencia para cada uno de los 5 algoritmos utilizando los parámetros predeterminados establecidos por Sklearn. Hay dos excepciones. La primera es incluir en todos los modelos el parámetro `random_state` con el objetivo de partir siempre de la misma semilla y controlar la aleatoriedad. La segunda, incorporar en el algoritmo de Regresión Logística el parámetro `iter_max = 150` con el fin de corregir el aviso: *STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.*

Posterior al entrenamiento, debemos de usar los datos de prueba para evaluar qué tan bien predicen los modelos. Para ello, utilizamos las siguientes 6 métricas:

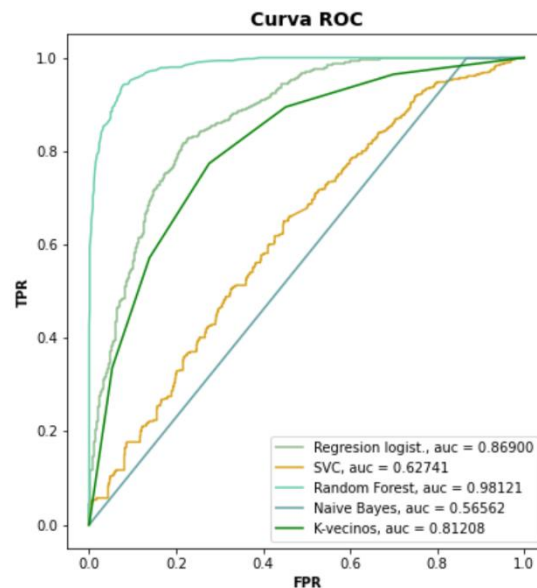
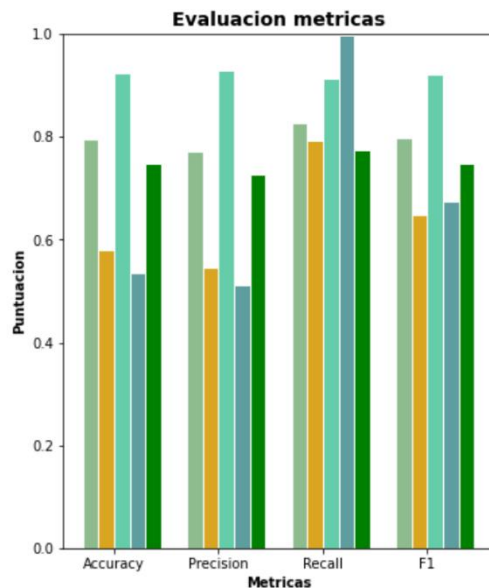
- Accuracy
- Precision
- Recall
- F1 score
- Área bajo la curva (AUC)
- Matriz de confusión

Como apoyo a la evaluación de cada modelo, creamos una función que calcula las 6 métricas mencionadas anteriormente. Tendrá como argumentos el modelo a evaluar y los conjuntos `X_test` e `y_test`.

La comparación de los resultados de rendimiento de cada modelo nos permitirá elegir entre todos los modelos candidatos cual es el que mejor se ajusta al problema planteado en este proyecto.

En nuestro caso, queremos centrarnos en predecir tantos positivos reales como podamos. Una clasificación errónea de un municipio con previsión de despoblarse en 10 años significar una oportunidad perdida para intentar cambiar la tendencia de dicho municipio.

Una vez entrenados y evaluados nuestros modelos, comparamos los resultados:



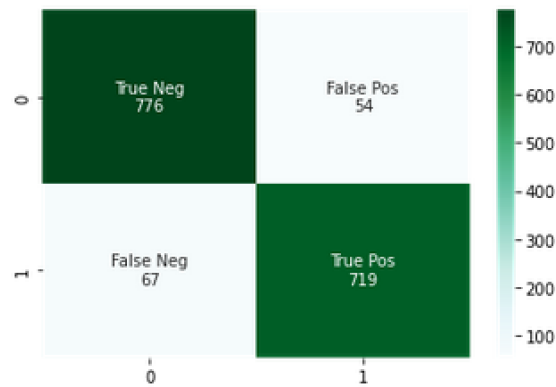
	Accuracy	Precision	Recall	F1
rl	0.796411	0.771056	0.826972	0.798036
svc	0.580446	0.547368	0.793893	0.647975
rf	0.925124	0.930142	0.914758	0.922386
nb	0.534653	0.511068	0.998728	0.676141
knn	0.748144	0.726404	0.773537	0.749230

Podemos observar que, de todos los modelos, Random Forest es el que mejores resultados obtiene en todas las métricas a excepción del Recall:

- **Accuracy:** con un 0,925 de predicciones correctas respecto del total de casos.
- **Precision:** 0,930 de porcentaje de casos positivos reales detectados.
- **Recall:** 0,914 de casos positivos que fueron correctamente identificadas. En Naive Bayes esta métrica es más precisa ya que detecta 66 casos más de TP que Random Forest.
- **F1:** con un valor de 0,922 esperado debido a los valores que toman la precision y el recall.

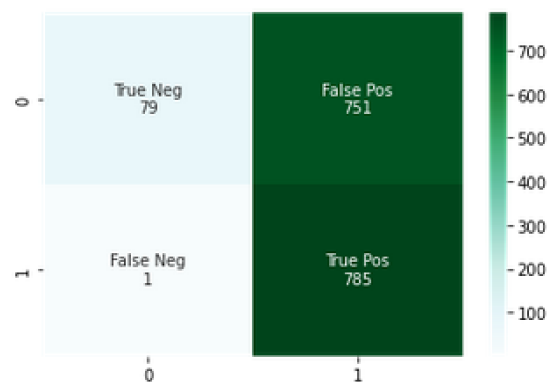
Si ahora nos centramos en la curva ROC, observamos que Random Forest es un modelo con una alta capacidad para discriminar adecuadamente entre despoblados y poblados. El área bajo la curva es de un 0,981, muy cerca de 1.

Como siguiente paso, para obtener una idea de cuán correctas fueron nuestras predicciones y cómo se comparan con los valores reales, observemos la matriz de confusión del modelo Random Forest:



El modelo clasifica como verdaderos positivos a 719 municipios, es decir, predice como despoblados los que efectivamente son despoblados. En cambio, hay 67 municipios que los clasifica como poblados cuando en realidad son despoblados.

Realizando como último paso una evaluación general, aunque el Recall de Naive Bayes es mejor que el de Random Forest, lo que nos asegura una menor pérdida de positivos, clasifica a 751 municipios como despoblados cuando son poblados, un alto número de falsas alarmas, lo que disminuye la precisión al 0,511.



Recopilando, podemos afirmar que Random Forest es el modelo más adecuado para resolver nuestro problema con muy buenos resultados y una alta precisión en el manejo de clasificación de las clases.

6.4 OPTIMIZACIÓN DEL MODELO

Aunque en el modelo base de Random Forest ya hemos obtenido un buen ajuste en las métricas, analicemos si estas son mejorables.

Para ello, vamos a ajustar los hiper parámetros que tenemos disponibles en el algoritmo, buscar sus valores óptimos, volver a evaluar el modelo y ver si tiene una mejoría en el rendimiento respecto de nuestro modelo base.

Para encontrar los valores óptimos de los hiper parámetros, iremos probando diferentes posibilidades. Para ello, utilizaremos la función GridSearchCV, disponible en Scikit-learn, que nos permite evaluar y seleccionar de forma fácil y sistemática los parámetros de un modelo utilizado la validación cruzada.

Después de múltiples pruebas, los hiper parámetros finalmente establecidos para determinar la configuración óptima de nuestro modelo son:

- **Max_depth:** [50, 60, 70]
- **Max_features:** [4, 6, 8, 9]
- **Min_samples_leaf:** [2, 3, 4, 5]
- **Min_samples_split:** [8, 10, 12]
- **N_estimators:** [100, 200, 300]

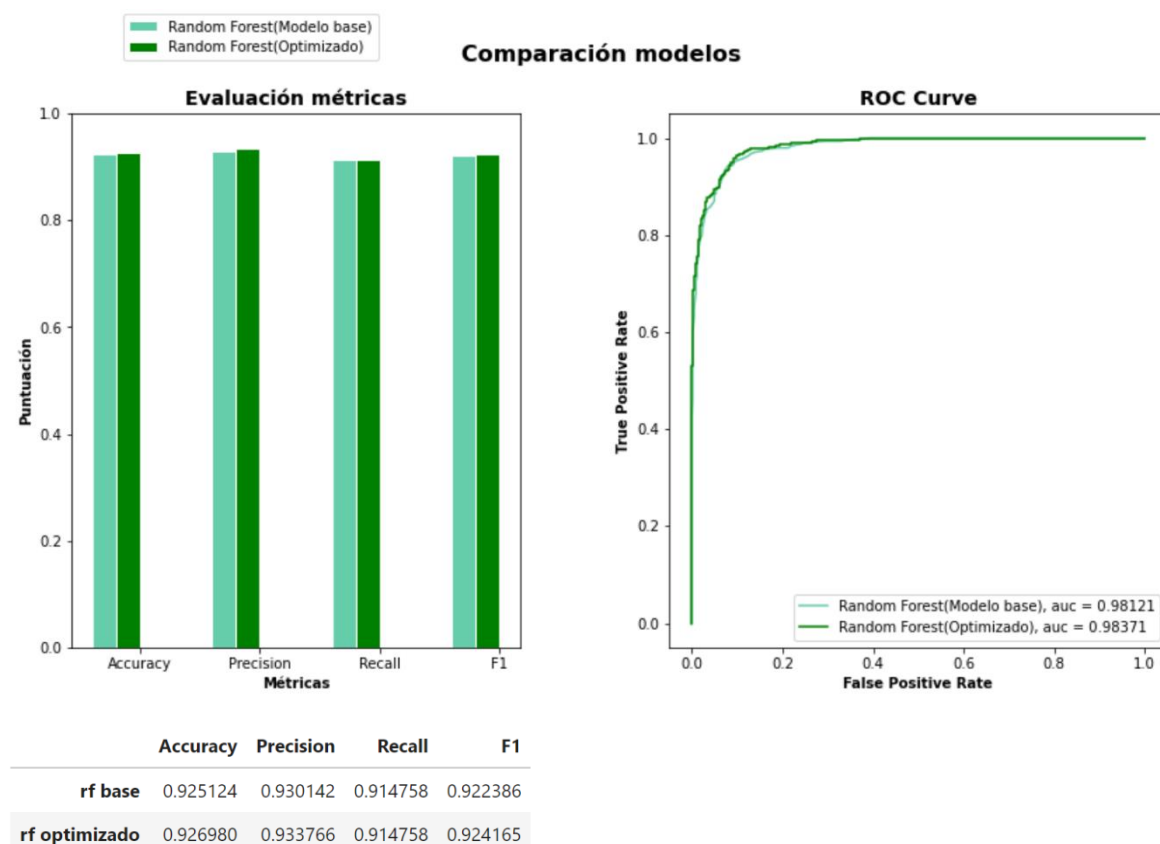
Además, la validación cruzada será de 5 divisiones, añadiendo adicionalmente `n_jobs = -1`, para acelerar el proceso de entrenamiento, y `verbose = 2` para visualizar el progreso del entrenamiento.

6.5 EVALUACIÓN DEL MODELO OPTIMIZADO

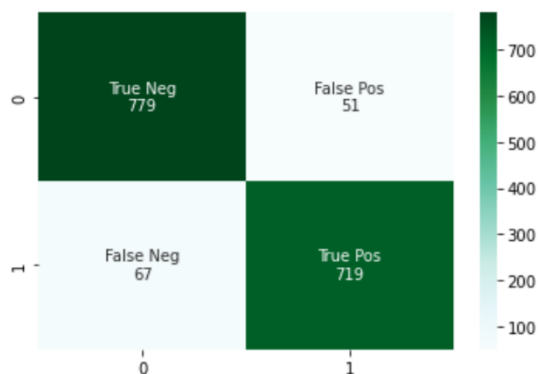
Los valores óptimos calculados por GridSearchCV son:

```
RandomForestClassifier(max_depth=50, max_features=9, min_samples_leaf=3,  
                        min_samples_split=8, n_estimators=200, random_state=0)
```

Una vez optimizado Random Forest, evaluamos el modelo y lo comparamos con el modelo base:



Aunque, como se ha comentado anteriormente, ya partimos de unos buenos resultados, vemos que el modelo optimizado funciona un poco mejor respecto del modelo base.

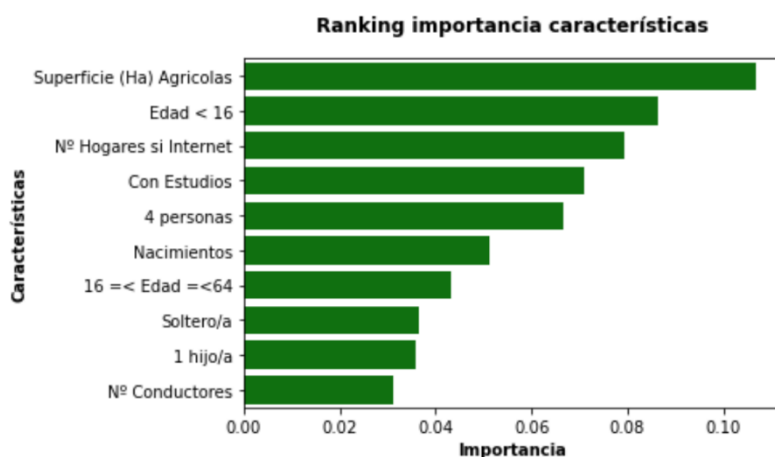


Si observamos la matriz de confusión del modelo optimizado, vemos que clasifica a 719 municipios como TP, al igual que nuestro modelo base, pero disminuye en 3 los municipios clasificados como FP.

Por tanto, podemos concluir que nuestro modelo optimizado tiene un rendimiento levemente superior al del modelo base.

6.6 IMPORTANCIA DE LAS CARACTERÍSTICAS

Como último paso, vamos a cuantificar la utilidad de las variables en el modelo Random Forest optimizado, observando la importancia relativa de cada una de las variables:



El gráfico muestra las 10 variables que más peso tienen en el ajuste y mejora del modelo.

Podemos observar, en la importancia devuelta, que no destaca ninguna variable significativamente sobre las demás, siendo los pesos de las variables valores no muy altos y con pocas diferencias entre ellas.

Por tanto, pesos bajos en las características, unido a que ya partimos de un buen ajuste del modelo, hace que descartemos realizar una selección de variables para volver a entrenar el modelo.

7. TABLEAU

Como último paso del proyecto, vamos a crear un libro de trabajo utilizando Tableau, con el que poder organizar todos los datos, interactuar sobre ellos, generar visualizaciones y compartir la información.

Una parte importante de este punto del proyecto, además de uno de los objetivos, es la centralización de los datos en una única ubicación. Hasta la fecha, hay que acudir a la fuente de origen de cada una de las variables para poder acceder a la información. Otro motivo relevante de esta unificación es la de visualizar de forma global, prácticamente todos los datos con los que se trabaja a lo largo del proyecto.

Como soporte de datos, se han utilizado los dos ficheros que contienen las entidades geométricas de la península y Canarias, y la última versión definitiva del fichero *Despoblacion*. Para la unión de estos archivos, se ha realizado un trabajo previo en el que se ha incorporado la variable *Natcode* al csv *Despoblacion*.

En el libro de trabajo, la información se organiza en los siguientes bloques/dashboards:

- Bloque demográfico
- Bloque económico
- Bloque geográfico
- Bloque hogar, vivienda y servicios
- Bloque despoblado y poblado

Se puede navegar a través de las diferentes vistas así como filtrar y visualizar la información por CCAA, provincias, municipios e incluso por municipios poblados y despoblados.



El reto de la despoblación a través del Data Science

<https://public.tableau.com/app/profile/adriana.borrajoviz/ElretodeladespoblacinatravsdeDataScience/Portada1>

Nota: ampliar a pantalla completa para una correcta visualización

Por último, comentar que un paso de mejora sería realizar la misma visualización pero con datos más actuales.