UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

Minería de datos y Modelización Predictiva

Datos Elecciones España

Juan Peñas Utrilla

Máster Big Data, Data Science & Inteligencia Artificial

Curso académico 2025-26

Índice

1. Introducción 3
2. Importación de los datos 3
3. Análisis descriptivo del conjunto de datos 4
4. Corrección de los errores detectados 5
5. Análisis de valores atípicos 6
6. Análisis de valores perdidos 8
7. Construcción del modelo de regresión lineal. 11
   7.1 Selección de variable clásica 11
   7.2 Selección del modelo ganador 14
   7.3 Interpretación de los coeficientes 15
8. Construcción del modelo de regresión logística. 17
   8.1 Selección de variable clásica 17
   8.2 Selección del modelo ganador 19
   8.3 Determinar el punto de corte óptimo 20
   8.4 Interpretación de los coeficientes 23
9. Conclusiones 25

1. Introducción

La participación electoral es uno de los indicadores fundamentales para evaluar la calidad democrática y el grado de implicación ciudadana en los procesos políticos. En este contexto, la abstención electoral no solo refleja desafección política, sino que suele estar estrechamente relacionada con factores socioeconómicos, demográficos y territoriales que condicionan el comportamiento electoral a nivel local. Comprender y anticipar estos patrones resulta clave tanto para el análisis político como para el diseño de políticas públicas orientadas a fomentar la participación ciudadana.

El presente trabajo tiene como objetivo analizar y modelizar el fenómeno de la abstención electoral a nivel municipal en España mediante técnicas de regresión. En particular, se abordan dos problemas complementarios: por un lado, la predicción del porcentaje de abstención en cada municipio, formulado como un problema de regresión lineal; y, por otro, la predicción de una variable dicotómica de abstención alta, que identifica aquellos municipios con niveles de abstención superiores a un umbral definido (30 %), planteado como un problema de regresión logística. Este enfoque dual permite capturar tanto la variabilidad continua del fenómeno como su clasificación en escenarios de riesgo elevado.

Para ello, se emplea un conjunto de datos que integra información demográfica y socioeconómica de los municipios españoles junto con los resultados electorales correspondientes. A partir de estas variables explicativas, se construyen modelos predictivos siguiendo un proceso riguroso de depuración de datos, análisis exploratorio y selección de variables, con el fin de garantizar la validez estadística, la interpretabilidad de los modelos y su capacidad de generalización.

El interés del estudio no se limita a la obtención de buenos resultados predictivos, sino también a la interpretación de los factores que influyen de manera significativa en la abstención electoral. De este modo, el trabajo combina el enfoque cuantitativo propio de la modelización estadística con una lectura analítica del comportamiento electoral, aportando una visión estructurada y fundamentada de los determinantes de la abstención a escala municipal.

2. Importación de los datos

Se comienza importando los datos, para ello se utiliza el siguiente código:

```
datos_original = pd.read_excel("DatosEleccionesEspaña.xlsx")
```

A continuación, es necesario eliminar aquellas variables que no se utilizarán para modelar, también se eliminan la variable ‘Name’ y ‘CodigoProvincia’, ya que no aportan información, y al no poderse transformar, tampoco se puede extraer valor de ellas:

```
columnas_eliminar = ['Izda_Pct', 'Dcha_Pct','Otros_Pct', 'Izquierda',
'Derecha','Name','CodigoProvincia']

datos = datos_original.drop(columns = columnas_eliminar)
```

Ahora se analiza que las variables hayan sido importadas correctamente con sus respectivos tipos, para ello se ejecuta `datos.dtypes`. Tras analizar la salida, es necesario modificar el tipo de la variable dicotómica objetivo:

```
AbstencionAlta                    int64
```

```
numericasAcategoricas = ['AbstencionAlta']

for var in numericasAcategoricas:

    datos[var] = datos[var].astype(str)
```

3. Análisis descriptivo del conjunto de datos

A continuación, se muestra un análisis descriptivo de las variables explicativas.

```
descriptivos_num = datos.describe().T

for num in numericas:

    descriptivos_num.loc[num, "Asimetria"] = datos[num].skew()

    descriptivos_num.loc[num, "Kurtosis"] = datos[num].kurtosis()

    descriptivos_num.loc[num, "Rango"] = np.ptp(datos[num].dropna().values)
```

Tabla 1. Análisis de rangos y detección de anomalías

Variable	min	max	Rango
Age_19_65_pct	23.459	100.002	76.543
Age_over65_pct	-18.052	76.472	94.524
Explotaciones	1	99999	99998
ForeignersPtge	-8.96	71.47	80.43
SameComAutonPtge	0	127.156	127.156

Se han detectado inconsistencias en ciertas variables porcentuales, las cuales presentan valores fuera del rango lógico (menores al 0 % o superiores al 100 %). Asimismo, se observan distribuciones con colas pesadas asociadas a las grandes ciudades; estos valores atípicos son legítimos y no deben considerarse errores. Por último, la variable 'Explotaciones' registra valores de 99999, lo que indica un código de valor perdido que requiere corrección.

Para detectar errores en las variables categóricas se muestran las frecuencias de cada una de sus categorías: `analizar_variables_categoricas(datos)`

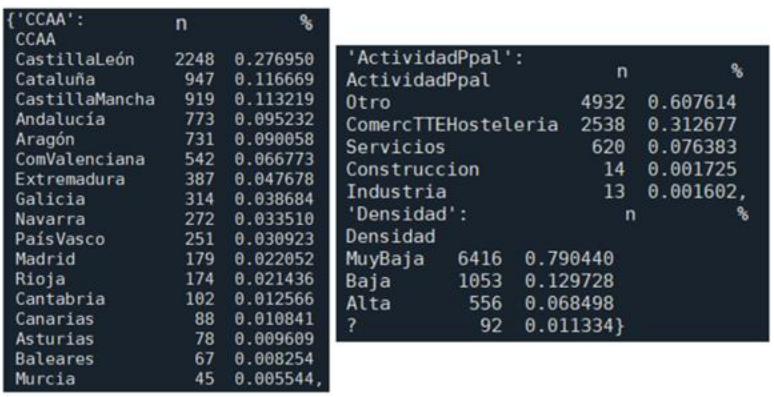

Fig. 1 Descriptivo de variables explicativas

Se ha detectado la presencia de caracteres anómalos (símbolo '?') en la variable ‘Densidad’, lo cual requiere una limpieza de formato. Adicionalmente, las variables ‘ActividadPpal’ y ‘CCAA’ muestran un desequilibrio en sus distribuciones, presentando categorías con una frecuencia marginal o poco representativa.

4. Corrección de los errores detectados

Se comienza corrigiendo las variables categóricas, primero se reagrupan las categorías de las variables ‘CCAA’ y ‘ActividadPpal’. Para las comunidades autónomas se agrupan aquellas que están representadas menos de 100 veces, agrupándose Canarias, Asturias, Baleares y Murcia.

```
# Conteo por CCAA

freq_ccaa = datos['CCAA'].value_counts()

# CCAA con menos de 100 observaciones

ccaa_poco_repr = freq_ccaa[freq_ccaa < 100].index

# Reagrupación

datos['CCAA'] = datos['CCAA'].replace(ccaa_poco_repr, 'Can_Ast_Bal_Mur')
```

Se procede de la misma manera para ‘ActividadPpal’ uniendo las categorías Servicios, Construccion e industria.

```
# Diccionario de recodificación

nueva_actividad = {

    'Servicios': 'Servicios_Constr_Industria',

    'Construccion': 'Servicios_Constr_Industria',

    'Industria': 'Servicios_Constr_Industria'}

# Reagrupación

datos['ActividadPpal'] = datos['ActividadPpal'].replace(nueva_actividad)
```

A continuación, se corrigen los valores 99999 de la variable ‘Explotaciones’:

```
datos['Explotaciones'] = datos['Explotaciones'].replace(99999, np.nan)
```

Se corrigen aquellas variables que representan porcentajes, para que todos los registros estén entre 0 y 100.

```
cols_ptge = datos.filter(regex=r'Ptge$').columns.tolist()

cols_ptge.append("Age_19_65_pct")

cols_ptge.append("Age_over65_pct")

for columna in cols_ptge:
```

```
    datos[columna] = [x if 0 <= x <= 100 else np.nan for x in datos[columna]]
```

Se cambian los valores de '?' en la variable 'Densidad' por nan:

```
datos['Densidad'] = datos['Densidad'].replace('?', np.nan)
```

5. Análisis de valores atípicos

Se indican los targets, ya que sobre estas variables no se calculan atípicos ni missings, y se separan del resto de variables explicativas.

```
varObjCont = datos['AbstentionPtge']

varObjBin = datos['AbstencionAlta']

datos_input = datos.drop(['AbstentionPtge', 'AbstencionAlta'], axis = 1)
```

Para evaluar la presencia de valores anómalos en las variables explicativas, se ha procedido a su detección mediante criterios estadísticos robustos, complementando este análisis con la visualización de la distribución de todas las variables en una escala normalizada [0, 1].

```
vars_numericas = datos_input.select_dtypes(include=[np.number]).columns

resultados_atipicos = []

for col in vars_numericas:

    # Se ha modificado atipicosAmissing, ahora la función devuelve: [serie_limpia,
numero_atipicos, valores_atipicos]

    _, n_atipicos, _ = atipicosAmissing(datos_input[col])

    # Guardamos el resultado

    resultados_atipicos.append({

        'Variable': col,

        'Num_Atipicos': n_atipicos,

        'Porcentaje': round((n_atipicos / len(datos_input)) * 100, 2)

    })

df_atipicos = pd.DataFrame(resultados_atipicos).sort_values('Num_Atipicos',
ascending=False)

datos_norm = datos_input[vars_numericas].copy()

datos_norm = (datos_norm - datos_norm.min()) / (datos_norm.max() - datos_norm.min())

datos_melted = datos_norm.melt(var_name='Variable', value_name='Valor Normalizado')

plt.figure(figsize=(10, 12))

sns.boxplot(
```

```
    data=datos_melted, x='Valor Normalizado', y='Variable', orient='h',
    palette='viridis', linewidth=1)
plt.title('Distribución de Variables y Valores Atípicos (Escala Normalizada 0-1)')
plt.xlabel('Rango Relativo')
plt.ylabel('Variables')
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
```

Tabla 2. Porcentaje de atípicos

Variable	Número de registros atípicos	Porcentaje del total de datos %
Servicios	981	12.09
totalEmpresas	851	10.48
Population	804	9.91
ComercTTEHosteleria	798	9.83
Pob2010	790	9.73

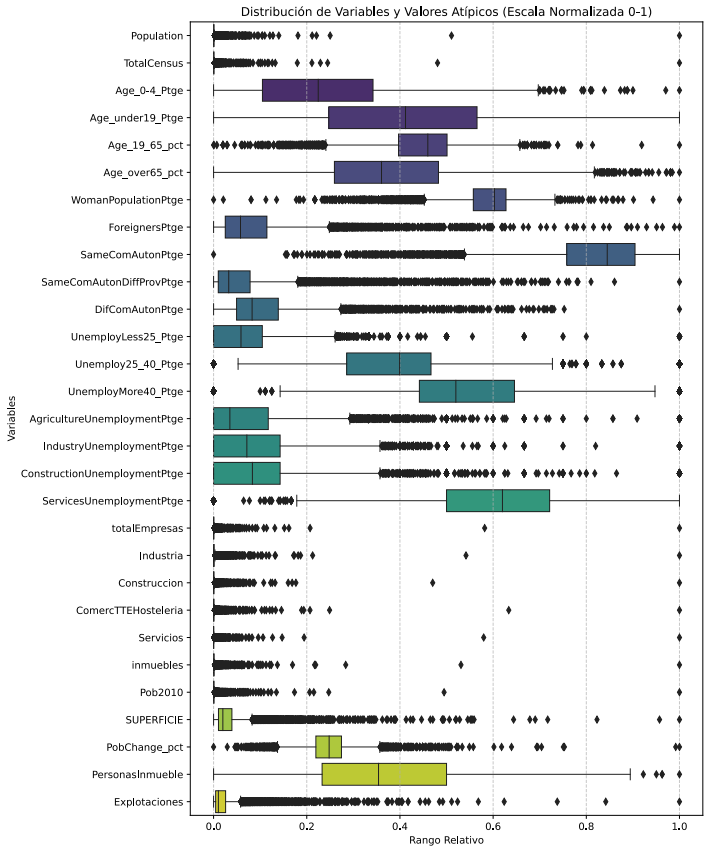

Fig. 2 Distribución de las variables explicativas en escala normalizada [0, 1] para la detección visual de valores atípicos y asimetrías.

Del análisis gráfico y numérico se desprende una clara dicotomía en el comportamiento de los datos:

- Variables poblacionales: Variables como 'Population', 'TotalCensus', 'totalEmpresas' o 'Servicios' acumulan el mayor porcentaje de atípicos (entre el 9% y el 12%). La distribución muestra una gran concentración de valores bajos (municipios pequeños) y una cola muy larga hacia la derecha, correspondiente a las capitales de provincia y grandes urbes.
- Variables porcentuales: Las variables porcentuales (ej. 'Age_over65_pct') presentan distribuciones más equilibradas y una tasa de atípicos nula o residual.

Tras el análisis exploratorio del conjunto de datos, se confirma la presencia de valores extremos en diversas variables, especialmente en aquellas asociadas al tamaño del municipio, como población, número de empresas o superficie. Sin embargo, estos valores no constituyen errores de medición ni observaciones anómalas desde el punto de vista conceptual, sino que reflejan la elevada heterogeneidad estructural existente entre municipios españoles, donde conviven pequeños núcleos rurales con grandes áreas urbanas.

Dado que estas variables representan magnitudes poblacionales bien definidas y no acotadas, todos sus valores son, en principio, posibles y coherentes con la realidad del fenómeno estudiado. En consecuencia, no se aplica ningún tratamiento de valores atípicos, ya que su eliminación o modificación podría introducir sesgos artificiales y conllevar una pérdida de información relevante. El proceso de depuración se centra exclusivamente en el tratamiento de valores perdidos y en la corrección de errores conceptuales previamente detectados, garantizando así la integridad estadística del conjunto de datos y la correcta interpretación posterior de los modelos de regresión.

6. Análisis de valores perdidos

Ahora se procede a imputar los valores perdidos, primero se analiza si hay correlaciones en el patrón de perdidos:

```
# Se ha modificado la función para mostrar correlaciones significativas

patron_perdidos(datos_input, threshold=0.01)
```

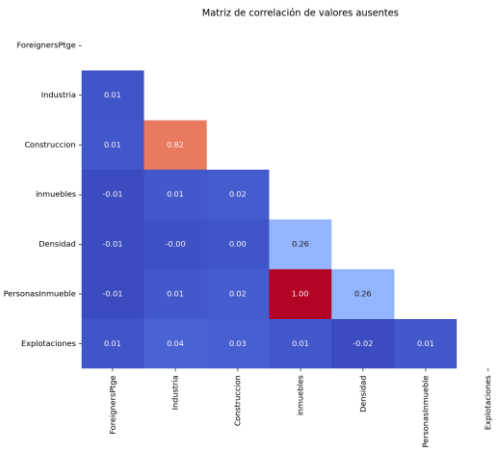

Fig. 3 Patrón de perdidos para aquellas variables con una correlación superior al 1 %

El análisis del patrón de valores ausentes muestra, en general, una baja correlación entre las ausencias de las distintas variables, lo que indica que los valores perdidos no siguen un patrón sistemático común. No obstante, se observa una correlación perfecta entre las variables 'Inmuebles' y 'PersonasInmueble'. En el resto de las variables, los valores ausentes se comportan de manera independiente, lo que permite abordar su tratamiento de forma individual sin introducir sesgos significativos. Para las variables correlacionadas se ha decidido imputar únicamente el número de inmuebles y recalcular posteriormente 'PersonasInmueble'.

Se calcule el porcentaje de perdidos por variable y se representa gráficamente:

```
variables_input = list(datos_input.columns)

prop_missingsVars = datos_input.isna().sum()/len(datos_input)

vars_con_missings = prop_missingsVars[prop_missingsVars >
0].sort_values(ascending=False)

vars_con_missings_pct = vars_con_missings * 100

plt.figure(figsize=(12, 6))

bars = plt.bar(vars_con_missings_pct.index, vars_con_missings_pct.values,

               color='blue', edgecolor='black', alpha=0.8)


plt.title('Porcentaje de Valores Perdidos por Variable', fontsize=14)

plt.ylabel('% de Nulos')

plt.xlabel('Variables')

plt.xticks(rotation=45, ha='right', fontsize=10)

plt.grid(axis='y', linestyle='--', alpha=0.5)

for bar in bars:

    height = bar.get_height()

    plt.text(bar.get_x() + bar.get_width()/2., height + 0.1,

             f'{height:.2f}%',

             ha='center', va='bottom', fontsize=9)

plt.show()
```

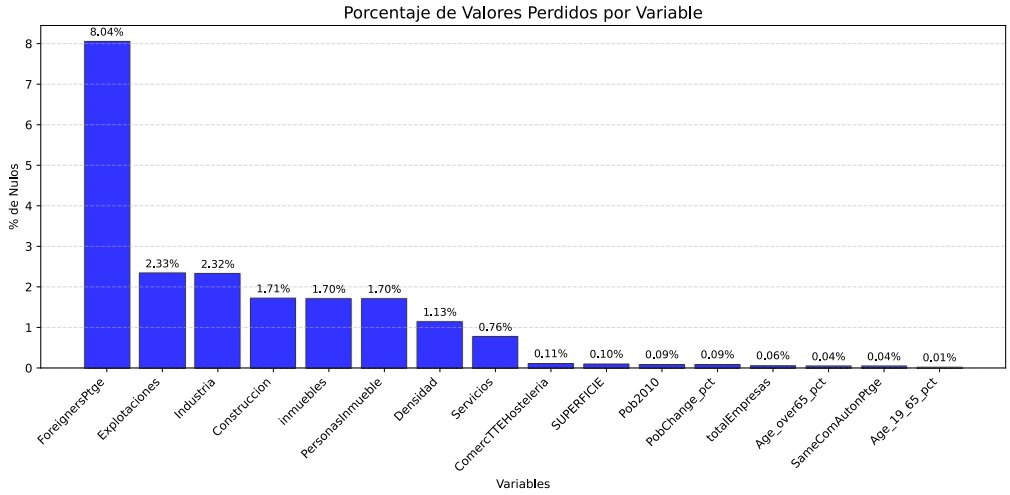

Fig. 4 Porcentaje de perdidos por variable, solo están representadas aquellas con más de un 0 %

Se observa que el porcentaje de valores perdidos es bajo en general, siendo la variable 'ForeignersPtge' la que presenta una mayor proporción de ausencias (8.04%). Dado que ninguna variable supera el umbral del 50% establecido para la eliminación de columnas, y que la eliminación de registros supondría una pérdida significativa de información muestral, se decide proceder a la imputación de datos.

Se revisa si hay alguna observación con más de un 50% de valores perdidos para eliminarla, sin embargo, tras ejecutar el filtro se comprueba que no hay ninguna fila que cumpla la condición.

```
missings_por_fila = datos_input.isna().sum(axis=1)

pct_missings_fila = (datos_input.isna().sum(axis=1) / datos_input.shape[1]) * 100

filas_a_eliminar = pct_missings_fila[pct_missings_fila > 50]
```

Debido a la existencia de valores atípicos y la asimetría observada en las variables numéricas, se utiliza la mediana como método de imputación robusto. Para las variables categóricas, se imputarán los valores ausentes utilizando la moda:

```
eliminar = [prop_missingsVars.index[x] for x in range(len(prop_missingsVars)) if
prop_missingsVars[x] > 0.5]

datos_input = datos_input.drop(eliminar, axis = 1)

for x in numericas_input:

    if x == "PersonasInmueble": # NO  imputar PersonasInmueble directamente

        continue

    datos_input[x] = ImputacionCuant(datos_input[x], 'mediana')

for x in categoricas_input:

    datos_input[x] = ImputacionCuali(datos_input[x], 'moda')

# Se recalcula PersonasInmueble tras imputar inmuebles

mask_recalc = datos_input["inmuebles"].notna() & (datos_input["inmuebles"] > 0)
```

```
datos_input.loc[mask_recalc, "PersonasInmueble"] = datos_input.loc[mask_recalc,
"Population"] / datos_input.loc[mask_recalc, "inmuebles"]
```

7. Construcción del modelo de regresión lineal.

7.1 Selección de variable clásica

En la construcción del modelo de regresión lineal se emplean métodos de selección clásica de variables, cuyo objetivo es identificar un subconjunto de variables explicativas que permita explicar adecuadamente la variabilidad de la variable objetivo, evitando modelos excesivamente complejos y reduciendo problemas como el sobreajuste o la multicolinealidad.

Estos métodos se basan en la comparación iterativa de modelos alternativos, evaluando en cada paso si la inclusión o exclusión de una variable mejora la calidad del modelo según un criterio de información. En este trabajo se utilizan tres estrategias clásicas de selección: forward, backward y stepwise, combinadas con los criterios AIC y BIC.

El método forward o hacia delante comienza con un modelo vacío, sin variables explicativas, e incorpora de forma iterativa aquella variable que produce la mayor mejora del modelo en cada paso. Una vez que una variable entra en el modelo, no puede ser eliminada en iteraciones posteriores.

Por el contrario, el método backward o hacia atrás parte de un modelo inicial que incluye todas las variables disponibles y elimina progresivamente aquellas que menos contribuyen a la explicación de la variable objetivo. Al igual que en el método forward, una vez que una variable es eliminada no puede volver a incorporarse.

El método stepwise combina ambos enfoques. En cada iteración se evalúa simultáneamente la posible entrada de nuevas variables y la eliminación de variables ya incluidas en el modelo, seleccionando la acción que mayor mejora produzca. Este enfoque es más flexible y permite corregir decisiones subóptimas tomadas en etapas anteriores del proceso.

La mejora o empeoramiento de los modelos se evalúa mediante dos criterios de información. El criterio de Akaike (AIC) prioriza el ajuste del modelo penalizando moderadamente su complejidad, mientras que el criterio bayesiano de Schwarz (BIC) impone una penalización más severa al número de parámetros, favoreciendo modelos más parsimoniosos. En general, BIC tiende a seleccionar modelos más simples que AIC.

Con el fin de comparar de manera sistemática estas estrategias, se construyen modelos utilizando las tres técnicas de selección (forward, backward y stepwise) bajo ambos criterios (AIC y BIC), empleando un conjunto de entrenamiento obtenido mediante una partición aleatoria de los datos. Posteriormente, los modelos resultantes se comparan atendiendo a medidas de ajuste y capacidad predictiva, seleccionando aquellos que presentan un mejor equilibrio entre complejidad y rendimiento.

A continuación, se muestra el código para realizar la partición de los datos input en train y test, y un poco más abajo el código para generar los modelos con selección clásica de variables y obtener parámetros representativos de los mismos.

```
x_train, x_test, y_train, y_test = train_test_split(datos_input,
np.ravel(varObjCont), test_size = 0.2, random_state = 123456)
```

```
var_cont1 = datos_input.select_dtypes(include=[np.number]).columns.tolist()

var_categ1 = datos_input.select_dtypes(exclude=[np.number]).columns.tolist()

modeloStepAIC = lm_stepwise(y_train, x_train, var_cont1, var_categ1, [], 'AIC')
modeloStepBIC = lm_stepwise(y_train, x_train, var_cont1, var_categ1, [], 'BIC')

modeloBackAIC = lm_backward(y_train, x_train, var_cont1, var_categ1, [], 'AIC')
modeloBackBIC = lm_backward(y_train, x_train, var_cont1, var_categ1, [], 'BIC')

modeloForwAIC = lm_forward(y_train, x_train, var_cont1, var_categ1, [], 'AIC')
modeloForwBIC = lm_forward(y_train, x_train, var_cont1, var_categ1, [], 'BIC')

modelos = {

    "Stepwise AIC": modeloStepAIC, "Stepwise BIC": modeloStepBIC,

    "Backward AIC": modeloBackAIC, "Backward BIC": modeloBackBIC,

    "Forward AIC":  modeloForwAIC, "Forward BIC":  modeloForwBIC,}

resumen_filas = []

for nombre, m in modelos.items():

    r2_train = Rsq(m['Modelo'], y_train, m['X'])

    x_test_m = crear_data_modelo(x_test, m['Variables']['cont'],
m['Variables']['categ'], [])

    r2_test  = Rsq(m['Modelo'], y_test, x_test_m)

    n_param = int(m['Modelo'].df_model + 1)

    resumen_filas.append({ "Modelo": nombre, "R2_Train": r2_train,

        "R2_Test": r2_test, "N_Param": n_param})

tabla_resultados = pd.DataFrame(resumen_filas).sort_values("R2_Test",ascending=False)
```

Tabla 3. Resultados de los modelos de regresión lineal

Método	R^2 Train	R^2 Test	Nº de Parámetros
Backward AIC	0.3800	0.3608	38
Backward BIC	0.3764	0.3547	31
Forward AIC	0.3671	0.3487	28
Stepwise AIC	0.3671	0.3487	28
Stepwise BIC	0.3642	0.3471	24
Forward BIC	0.3642	0.3471	24

También se procede a realizar validación cruzada con los modelos generados, con el objetivo de evaluar su robustez y capacidad de generalización. Este procedimiento consiste en dividir el conjunto de datos de entrenamiento en *k* subconjuntos o *folds* de tamaño similar. En cada iteración, uno de estos subconjuntos se utiliza como conjunto de validación, mientras que los *k−1* restantes se emplean para ajustar el modelo. El proceso se repite hasta que todos los subconjuntos

han actuado una vez como validación, lo que permite obtener una estimación del rendimiento del modelo menos dependiente de una única partición de los datos. De este modo, la validación cruzada proporciona información tanto sobre el rendimiento medio como sobre su variabilidad, facilitando la comparación entre modelos y reduciendo el riesgo de sobreajuste.

```
# Crea un DataFrame vacío para almacenar resultados

results = pd.DataFrame({‘Rsquared’: [], ‘Resample’: [], ‘Modelo’: []})

# Realiza el proceso 20 veces

for rep in range(20):

    for nombre, m in modelos.items():

        # Validacion cruzada de 5 folds para cada modelo

        cv_scores =  egression_cruzada_lm(5, x_train, y_train,

                                          m[‘Variables’][‘cont’],

                                          m[‘Variables’][‘categ’], [])

        # Almacenar resultados

        temp_df = pd.DataFrame({ ‘Rsquared’: cv_scores,

            ‘Resample’: [‘Rep’ + str(rep + 1)] * 5,

            ‘Modelo’: [nombre] * 5 })

        results = pd.concat([results, temp_df], axis=0)


plt.figure(figsize=(10, 6))

sns.boxplot(data=results, x=’Modelo’, y=’Rsquared’, showfliers=False)

sns.stripplot(data=results, x=’Modelo’, y=’Rsquared’, jitter=False, color=’black’,
alpha=0.5)

plt.title(“Distribución del R² por modelo (CV Repetida)”)

plt.xticks(rotation=45)

plt.grid(axis=’y’, alpha=0.3)

plt.tight_layout()

plt.savefig(‘boxplot_regresion_lineal.svg’)

plt.show()
```

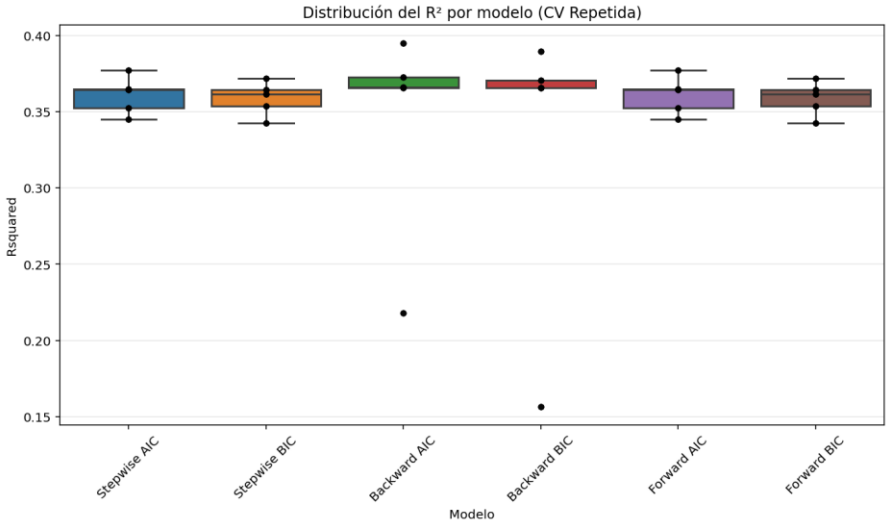

Fig. 5 Resultados de la validación cruzada para regresión lineal

La Tabla 3 recoge los resultados obtenidos para los distintos modelos de regresión lineal construidos mediante métodos de selección clásica, combinando las estrategias forward, backward y stepwise con los criterios de información AIC y BIC. Para cada modelo se presenta el coeficiente de determinación en el conjunto de entrenamiento y en el conjunto de test, así como el número de parámetros estimados, lo que permite evaluar conjuntamente el rendimiento predictivo y la complejidad de cada alternativa.

7.2 Selección del modelo ganador

Del análisis comparativo se observa que el modelo Backward AIC alcanza el mayor poder predictivo en el conjunto de test, con un valor de $R^2 = 0.3608$. No obstante, esta mejora en capacidad explicativa es relativamente limitada y se obtiene a costa de una complejidad elevada, ya que el modelo requiere la estimación de 39 parámetros, lo que incrementa el riesgo de sobreajuste y dificulta la interpretación de los resultados.

Por el contrario, las estrategias basadas en el criterio BIC, tanto Stepwise como Forward, convergen hacia soluciones significativamente más parsimoniosas, con 25 parámetros. Aunque su coeficiente de determinación en test es ligeramente inferior ($R^2 = 0.3471$), la pérdida de capacidad explicativa respecto al modelo más complejo es marginal (en torno a un 1.37 %), mientras que la reducción en complejidad es sustancial, eliminando 14 variables del modelo.

La Figura 5, que muestra la distribución del R^2 obtenida mediante validación cruzada repetida, refuerza esta conclusión. Se observa que todos los modelos presentan distribuciones muy similares, con medianas próximas y una variabilidad comparable entre particiones. En particular, los modelos basados en BIC no muestran un deterioro significativo en estabilidad predictiva frente a los modelos más complejos, lo que indica que las diferencias observadas en una única partición train/test no son estructurales.

Por estos motivos, se selecciona como modelo final el obtenido mediante selección stepwise con criterio BIC. Esta elección se fundamenta en su carácter parsimonioso, su estabilidad predictiva evidenciada por la validación cruzada y su menor riesgo de sobreajuste, lo que lo convierte en una opción más robusta y adecuada para la interpretación y el análisis posterior de los resultados.

El modelo final de regresión lineal presenta un coeficiente de determinación en entrenamiento de

$R^2 = 0.364$, lo que indica que el conjunto de variables seleccionadas explica aproximadamente un 36 % de la variabilidad del porcentaje de abstención municipal. Este nivel de ajuste es razonable teniendo en cuenta la naturaleza social del fenómeno analizado, caracterizado por una elevada heterogeneidad y la influencia de factores no observables.

```
modeloStepBIC['Modelo'].summary()
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.364
Model:                            OLS   Adj. R-squared:                  0.362
Method:                 Least Squares   F-statistic:                     161.2
Date:                Thu, 15 Jan 2026   Prob (F-statistic):               0.00
Time:                        16:08:29   Log-Likelihood:                -20896.
No. Observations:                6493   AIC:                         4.184e+04
Df Residuals:                    6469   BIC:                         4.200e+04
Df Model:                          23                                         
Covariance Type:            nonrobust                                         
==============================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                         34.0243      1.171     29.043      0.000      31.728      36.321
SameComAutonPtge              -0.0581      0.007     -7.805      0.000      -0.073      -0.044
ServicesUnemploymentPtge       0.0240      0.003      6.861      0.000       0.017       0.031
ConstructionUnemploymentPtge   0.0409      0.006      6.835      0.000       0.029       0.053
Explotaciones                  0.0019      0.000      4.749      0.000       0.001       0.003
Age_19_65_pct                 -0.0646      0.013     -4.948      0.000      -0.090      -0.039
SUPERFICIE                 3.471e-05   8.77e-06      3.957      0.000    1.75e-05    5.19e-05
Age_under19_Ptge               0.0695      0.018      3.918      0.000       0.035       0.104
IndustryUnemploymentPtge       0.0254      0.007      3.841      0.000       0.012       0.038
CCAA_Aragón                   -1.7665      0.393     -4.498      0.000      -2.536      -0.997
CCAA_Can_Ast_Bal_Mur           3.7637      0.491      7.669      0.000       2.802       4.726
CCAA_Cantabria                -1.0543      0.690     -1.528      0.127      -2.407       0.299
CCAA_CastillaLeón             -2.4038      0.336     -7.159      0.000      -3.062      -1.746
CCAA_CastillaMancha           -5.3507      0.375    -14.281      0.000      -6.085      -4.616
CCAA_Cataluña                  6.3450      0.365     17.380      0.000       5.629       7.061
CCAA_ComValenciana            -6.7876      0.407    -16.659      0.000      -7.586      -5.989
CCAA_Extremadura              -1.1396      0.434     -2.629      0.009      -1.990      -0.290
CCAA_Galicia                   2.0339      0.482      4.222      0.000       1.090       2.978
CCAA_Madrid                   -3.9466      0.598     -6.594      0.000      -5.120      -2.773
CCAA_Navarra                   2.8403      0.504      5.635      0.000       1.852       3.828
CCAA_PaísVasco                 4.0207      0.507      7.923      0.000       3.026       5.016
CCAA_Rioja                    -7.2815      0.615    -11.843      0.000      -8.487      -6.076
ActividadPpal_Otro            -2.3341      0.231    -10.091      0.000      -2.787      -1.881
ActividadPpal_Servicios_Constr_Industria    -1.0463      0.319     -3.285      0.001      -1.671      -0.422
==============================================================================
Omnibus:                      227.442   Durbin-Watson:                   2.030
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              511.512
Skew:                           0.205   Prob(JB):                    8.45e-112
Kurtosis:                       4.313   Cond. No.                     1.97e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.97e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Fig. 6 Summary del modelo ganador con principales estadísticos

El contraste global del modelo confirma su significatividad estadística. En particular, el estadístico $F = 161.12$ con un *p-valor* inferior a 0.001 permite rechazar la hipótesis nula de que todos los coeficientes sean simultáneamente nulos, indicando que el conjunto de variables incluidas aporta información relevante frente a un modelo sin predictores. Estos resultados respaldan la validez global del modelo como herramienta explicativa del comportamiento de la abstención electoral.

7.3 Interpretación de los coeficientes

A continuación, se interpreta el efecto de dos variables significativas incluidas en el modelo ganador (Stepwise BIC), una de naturaleza continua y otra categórica, manteniendo constantes el resto de factores.

Variable continua: ‘ServicesUnemploymentPtge’:

El coeficiente estimado para la tasa de desempleo en el sector servicios es positivo y estadísticamente significativo ($\beta = 0.0240$, $p < 0.001$). Esto implica que, manteniendo constantes el resto de variables del modelo, un incremento de un punto porcentual en la tasa de desempleo del sector servicios se asocia con un aumento medio de 0.024 puntos porcentuales en la abstención electoral municipal. Este resultado es coherente con la literatura sobre comportamiento electoral, que vincula situaciones de precariedad laboral y desempleo con mayores niveles de desafección política y una menor participación en los procesos electorales.

Variable categórica: ‘CCAA_Cataluña’:

La variable dummy correspondiente a Cataluña presenta un coeficiente positivo de elevada magnitud ($\beta = 6.3450$) y altamente significativo ($p < 0.001$) en comparación con la categoría de referencia (Andalucía). Controlando por el resto de variables demográficas y socioeconómicas incluidas en el modelo, los municipios situados en Cataluña muestran, en promedio, una tasa de abstención aproximadamente 6.34 puntos porcentuales superior a la de los municipios de la categoría de referencia. Este efecto fijo regional sugiere la existencia de factores territoriales no observables (de carácter político, institucional o cultural) que influyen diferencialmente en la movilización del electorado y que no quedan completamente explicados por las variables estructurales consideradas.

```
# Calcular la importancia de variables para el modelo lineal ganador

importancia_lineal = modelEffectSizes(modeloStepBIC['Modelo'], y_train, x_train,


modeloStepBIC['Variables']['cont'],modeloStepBIC['Variables']['categ'])
```

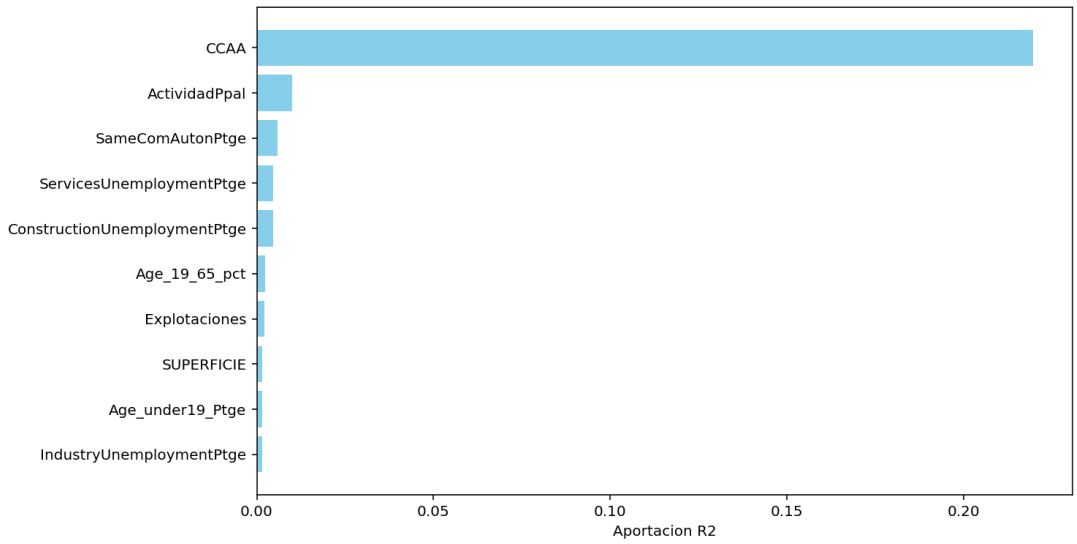

Fig. 7 Importancia de las variables del modelo lineal ganador

La Figura 7 ilustra la importancia relativa de las variables incluidas en el modelo lineal final (Stepwise BIC), cuantificada mediante su aportación marginal al coeficiente de determinación R^2. Este análisis permite jerarquizar los predictores según su capacidad real para explicar la variabilidad del porcentaje de abstención.

Del gráfico se desprende una hegemonía indiscutible del factor territorial. La variable CCAA presenta una aportación al R^2 superior a 0.20, una magnitud desproporcionadamente mayor que la del resto de variables explicativas combinadas. Esto confirma que el porcentaje de abstención es un fenómeno fuertemente regionalizado, donde la ubicación geográfica del municipio absorbe la mayor parte de la capacidad predictiva del modelo, muy por encima de las características demográficas locales.

En un segundo nivel de importancia, aunque con una contribución mucho más modesta, aparecen variables estructurales como la actividad principal económica (ActividadPpal) y el arraigo poblacional (SameComAutonPtge). Esto sugiere que el perfil productivo del municipio y la estabilidad residencial de sus habitantes aportan matices relevantes al modelo, una vez descontado el efecto regional.

Por último, resulta llamativo observar que las variables puramente económicas, como las tasas de desempleo sectoriales (ConstructionUnemploymentPtge y ServicesUnemploymentPtge), tienen una aportación marginal a la varianza explicada (aportación al R^2 inferior a 0.02). Aunque anteriormente vimos que sus coeficientes eran estadísticamente significativos, este gráfico revela que su "fuerza" explicativa es residual comparada con el peso determinante de la Comunidad Autónoma.

8. Construcción del modelo de regresión logística.

8.1 Selección de variable clásica

En este apartado se aborda la modelización de la variable binaria *AbstencionAlta*, definida como aquella que toma el valor 1 cuando el porcentaje de abstención supera el 30 % y 0 en caso contrario. Dado el carácter dicotómico de la variable objetivo, el problema se formula mediante un modelo de regresión logística, que permite estimar la probabilidad de que un municipio presente niveles elevados de abstención electoral en función de sus características demográficas y socioeconómicas.

Los datos han sido previamente depurados y tratados, por lo que el proceso de modelización se inicia directamente a partir del conjunto de variables explicativas limpio. El conjunto de datos se divide en subconjuntos de entrenamiento y test, empleando un 80 % de las observaciones para el ajuste del modelo y el 20 % restante para su evaluación, garantizando la reproducibilidad mediante una semilla fija. El mismo procedimiento para la regresión lineal.

Para la selección de variables se emplean los métodos clásicos forward, backward y stepwise, combinados con los criterios de información AIC y BIC. Estos métodos permiten comparar modelos alternativos atendiendo al compromiso entre capacidad explicativa y complejidad, evitando modelos innecesariamente sobreajustados.

Se construyen un total de seis modelos, cuyos resultados se comparan en términos de AUC en test, AUC media en validación cruzada y número de parámetros estimados.

```
x_train, x_test, y_train, y_test = train_test_split(datos_input, np.ravel(varObjBin),
test_size = 0.2, random_state = 13)

y_train, y_test = y_train.astype(int), y_test.astype(int)

var_cont = x_train.select_dtypes(include=[np.number]).columns.tolist()
```

```
var_categ = x_train.select_dtypes(exclude=[np.number]).columns.tolist()

mStepAIC = glm_stepwise(y_train, x_train, var_cont, var_categ, [], 'AIC')

mStepBIC = glm_stepwise(y_train, x_train, var_cont, var_categ, [], 'BIC')

mBackAIC = glm_backward(y_train, x_train, var_cont, var_categ, [], 'AIC')

mBackBIC = glm_backward(y_train, x_train, var_cont, var_categ, [], 'BIC')

mForwAIC = glm_forward(y_train, x_train, var_cont, var_categ, [], 'AIC')

mForwBIC = glm_forward(y_train, x_train, var_cont, var_categ, [], 'BIC')

# Tabla comparativa (AUC test, AUC CV, pseudoR2, nº parámetros)

modelos = {

    "Stepwise AIC": mStepAIC, "Stepwise BIC": mStepBIC,

    "Backward AIC": mBackAIC, "Backward BIC": mBackBIC,

    "Forward AIC":  mForwAIC, "Forward BIC":  mForwBIC,}

K = 5
REPS = 20

results_auc = []   # para boxplot (todas las AUC de cada fold y repetición)

filas = []         # para tabla resumen

for nombre, m in modelos.items():

    x_test_m = crear_data_modelo(x_test, m['Variables']['cont'],
m['Variables']['categ'], [])

    auc_test = roc_auc_score(y_test, m['Modelo'].predict_proba(x_test_m)[:, 1])

    aucs_all = []

    for rep in range(REPS):

        aucs = validacion_cruzada_glm(K, x_train, y_train,
m['Variables']['cont'],m['Variables']['categ'], [])

        for auc in aucs:

            results_auc.append({"Modelo": nombre, "AUC_CV": auc, "Rep": rep + 1})

        aucs_all.extend(aucs)

    auc_cv_mean = float(np.mean(aucs_all))

    pr2_train = pseudoR2(m['Modelo'], m['X'], y_train)

    n_param = len(m['Modelo'].coef_[0]) + 1

    filas.append([nombre, auc_test, auc_cv_mean, pr2_train, n_param])
```

```
tabla = (pd.DataFrame(filas, columns=["Modelo", "AUC_Test", "AUC_CV",
"PseudoR2_Train", "N_param"]).sort_values(["AUC_CV", "AUC_Test"], ascending=False))

df_auc = pd.DataFrame(results_auc)
```

Tabla 4. Resultados de los modelos de regresión logística

Modelo	AUC Test	AUC CV	pseudo-R^2 Train	Nº de Parámetros
Backward AIC	0.807	0.822	0.265	28
Backward BIC	0.807	0.822	0.265	28
Forward AIC	0.809	0.817	0.257	25
Forward BIC	0.809	0.817	0.257	25
Stepwise AIC	0.806	0.817	0.256	23
Stepwise BIC	0.806	0.817	0.256	23

La Tabla 4 resume las métricas obtenidas para los distintos modelos de regresión logística construidos mediante los métodos de selección clásica (forward, backward y stepwise) combinados con los criterios de información AIC y BIC. Para cada modelo se presenta el valor del AUC en el conjunto de test, el AUC medio obtenido mediante validación cruzada, el pseudo-R^2 en entrenamiento y el número de parámetros estimados, lo que permite evaluar simultáneamente capacidad discriminante, robustez y complejidad.

8.2 Selección del modelo ganador

Del análisis comparativo se observa que todos los modelos presentan un rendimiento predictivo muy similar, con valores de AUC en test comprendidos en un rango muy estrecho (entre 0.806 y 0.809). En particular, los modelos Backward (AIC y BIC) alcanzan el máximo AUC en test (0.807) y el mayor pseudo-R^2en entrenamiento (0.265)

Antes de tomar una decisión, es fundamental contextualizar estos valores de pseudo-R^2. Aunque numéricamente parecen bajos respecto a un R^2 lineal, en regresión logística valores entre 0.2 y 0.4 se consideran indicativos de un ajuste excelente (equivalente a un R^2de 0.7-0.9 en OLS). Dado que todos los modelos superan holgadamente el umbral de 0.2 (oscilando entre 0.256 y 0.265), podemos afirmar que todos ellos capturan la estructura de los datos de forma muy satisfactoria.

Al comparar estas métricas con la complejidad de los modelos, se observa que la estrategia Backward requiere estimar 28 parámetros para lograr ese ligero incremento en ajuste. Por el contrario, la estrategia Stepwise consigue un rendimiento prácticamente idéntico (AUC de 0.806 y pseudo-R^2 de 0.256) empleando solo 23 parámetros. Esta diferencia marginal en la capacidad explicativa (apenas 0.009 puntos de pseudo-R^2) no justifica la inclusión de cinco variables adicionales, lo que inclina la balanza hacia los modelos Stepwise por su mayor parsimonia y facilidad de interpretación.

Con el fin de evaluar la estabilidad y capacidad de generalización de los modelos, se representan los resultados de la validación cruzada k-fold, utilizando el AUC como métrica de evaluación.

```
# Boxplot de AUC-CV por modelo (robustez)

plt.figure(figsize=(10,5))

sns.boxplot(data=df_auc, x="Modelo", y="AUC_CV", showfliers=True)
```

```
sns.stripplot(data=df_auc, x="Modelo", y="AUC_CV", color="black", alpha=0.6,
jitter=False)

plt.xticks(rotation=30)

plt.grid(axis="y", alpha=0.3)

plt.title("Distribución de AUC por modelo (CV)")

plt.savefig("grafico_modelos_logistica.svg", bbox_inches="tight")

plt.show()
```

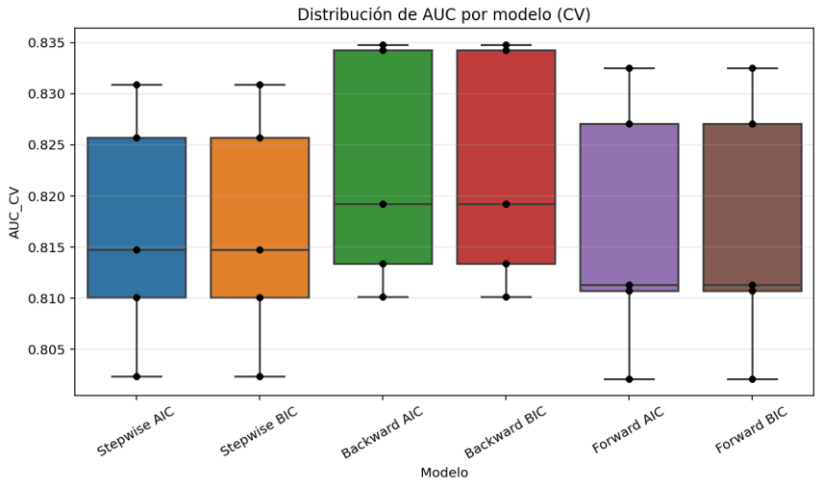

Fig. 8 Resultados de la validación cruzada para regresión logística

Los resultados se representan mediante diagramas de cajas, tal como se muestra en la Fig 8. El análisis gráfico revela distribuciones de AUC muy próximas entre modelos, con medianas similares y rangos intercuartílicos comparables, lo que indica que el rendimiento es estable y poco dependiente de una partición concreta de los datos. No se observan diferencias sistemáticas que justifiquen la elección de un modelo claramente más complejo desde el punto de vista predictivo.

Atendiendo conjuntamente a los resultados en test, a la validación cruzada y a la complejidad de los modelos, se selecciona como modelo final el obtenido mediante selección Stepwise con criterio BIC. Este modelo presenta un valor de AUC en test (AUC ≈ 0.806) prácticamente equivalente al de las alternativas con mayor rendimiento, pero con un menor número de parámetros estimados, lo que reduce el riesgo de sobreajuste y mejora la interpretabilidad del modelo.

La capacidad discriminante del modelo seleccionado se evalúa posteriormente mediante la curva ROC obtenida sobre el conjunto de test.

8.3 Determinar el punto de corte óptimo

Ahora se obtiene la curva ROC del modelo Stepwise BIC:

```
modelo_ganador = mStepBIC

var_cont_g  = modelo_ganador['Variables']['cont']

var_categ_g = modelo_ganador['Variables']['categ']
```

```
var_inter_g = modelo_ganador['Variables'].get('inter', [])

x_test_m = crear_data_modelo(x_test, var_cont_g, var_categ_g, var_inter_g)

auc_test = curva_roc(x_test_m, y_test, modelo_ganador)

>> Área bajo la curva ROC = 0.8064023746241433
```

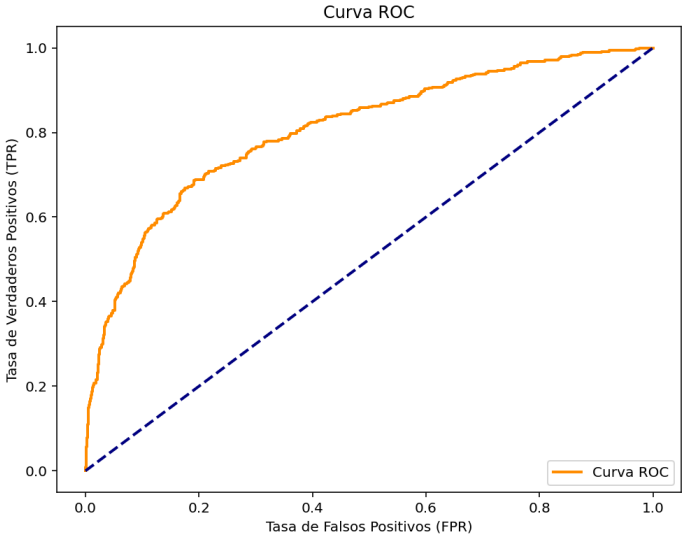

Fig. 9 Curva ROC del modelo ganador

En la figura 9 se aprecia la curva ROC cuyo área bajo la curva es de 0.8, siendo este un buen resultado para un primer estudio. Dado que la regresión logística proporciona probabilidades estimadas, es necesario definir un punto de corte para la clasificación final. Este se determina maximizando el índice de Youden, que equilibra simultáneamente la sensibilidad y la especificidad del modelo.

```
grid = np.linspace(0.01, 0.99, 99)

mejor_p, mejor_J, mejor_met = None, -999, None

for p in grid:

    met = sensEspCorte(modelo_ganador['Modelo'], x_test, y_test, float(p),

                       var_cont_g, var_categ_g, var_inter_g)  # devuelve 1 fila

    sens = float(met["Sensitivity"].iloc[0])

    spec = float(met["Specificity"].iloc[0])

    J = sens + spec - 1

    if J > mejor_J:

        mejor_J, mejor_p, mejor_met = J, float(p), met

# Matriz de confusión con ese corte

prob = modelo_ganador['Modelo'].predict_proba(x_test_m)[:,1]
```

```
yhat = (prob > mejor_p).astype(int)

tn, fp, fn, tp = confusion_matrix(y_test, yhat).ravel()

pd.DataFrame([[tn, fp],[fn, tp]], index=["Real 0","Real 1"], columns=["Pred 0","Pred 1"])
```

Tabla 5. Punto de corte y matriz de confusión del modelo ganador

Punto corte: 0.32	**Pred 0**	**Pred 1**
Real 0	925	226
Real 1	147	326

Dado que el modelo de regresión logística proporciona probabilidades estimadas, es necesario definir un punto de corte para convertir dichas probabilidades en una clasificación binaria. Este umbral se determina maximizando el índice de Youden, definido como $J = \text{Sensibilidad} + \text{Especificidad} - 1$, ya que permite equilibrar simultáneamente la capacidad del modelo para detectar correctamente la clase positiva y la clase negativa.

El punto de corte óptimo obtenido es 0.32, inferior al valor convencional de 0.5. Este resultado es coherente con la distribución de la variable objetivo y refleja la necesidad de reducir el número de falsos negativos en la detección de municipios con alta abstención electoral, aceptando a cambio un mayor número de falsos positivos.

A partir de estos resultados, se obtiene una sensibilidad aproximada del 68.9 %, lo que indica que el modelo identifica correctamente cerca de tres cuartas partes de los municipios con alta abstención, y una especificidad del 80.4 %, reflejando una buena capacidad para descartar municipios con niveles bajos de abstención. Este equilibrio entre sensibilidad y especificidad confirma que el punto de corte seleccionado es adecuado y coherente con el objetivo del análisis.

En conjunto, el análisis del punto de corte y de la matriz de confusión muestra que el modelo presenta una capacidad de clasificación razonable, reproduciendo de forma equilibrada la estructura real de la variable binaria y evitando sesgos sistemáticos en favor de una de las clases.

```
X_modelo_float = modelo_ganador['X'].astype(float)

resumen_log = summary_glm(modelo_ganador['Modelo'], y_train, X_modelo_float)

coef = resumen_log['Contrastes'].copy()

coef['Odds Ratio'] = np.exp(coef['Estimate'])

coef

resumen_log['BondadAjuste']
```

```
In [29]: coef
Out[29]:
                                    Variable  Estimate  ...  signif  Odds Ratio
0                                (Intercept)  4.132598  ...     ***   62.339670
1                             SameComAutonPtge -0.025844  ...     ***    0.974487
2                                Age_19_65_pct -0.028930  ...     ***    0.971484
3                   ConstructionUnemploymentPtge  0.009784  ...     ***    1.009832
4                                  SUPERFICIE  0.000014  ...     ***    1.000014
5                    AgricultureUnemploymentPtge -0.007335  ...       *    0.992692
6                               CCAA_Aragón -0.646888  ...     ***    0.523673
7                         CCAA_Can_Ast_Bal_Mur  0.891841  ...     ***    2.439617
8                               CCAA_Cantabria -1.397254  ...     ***    0.247275
9                            CCAA_CastillaLeón -0.801769  ...     ***    0.448535
10                          CCAA_CastillaMancha -1.572428  ...     ***    0.207541
11                               CCAA_Cataluña  2.094785  ...     ***    8.123696
12                           CCAA_ComValenciana -2.660047  ...     ***    0.069945
13                              CCAA_Extremadura -0.377928  ...       *    0.685280
14                                  CCAA_Galicia  0.453781  ...      **    1.574254
15                                  CCAA_Madrid -1.642171  ...     ***    0.193559
16                                 CCAA_Navarra  0.594160  ...     ***    1.811508
17                              CCAA_PaísVasco  0.803262  ...     ***    2.232813
18                                   CCAA_Rioja -2.374331  ...     ***    0.093077
19                           ActividadPpal_Otro -0.733632  ...     ***    0.480162
20  ActividadPpal_Servicios_Constr_Industria -0.581574  ...     ***    0.559018
21                              Densidad_Baja -0.260952  ...       .    0.770318
22                           Densidad_MuyBaja -0.705886  ...     ***    0.493671

Out[28]:
           LLK          AIC          BIC
0 -3015.414877  6034.829754  6048.386714
```

Fig. 10 Summary del modelo ganador de regresión logística

La Figura 10 presenta el resumen del modelo final de regresión logística, seleccionado mediante el procedimiento Stepwise con criterio BIC. En ella se muestran los coeficientes estimados, su significatividad estadística y las métricas globales de ajuste, lo que permite evaluar tanto la calidad del modelo como la contribución individual de las variables explicativas.

Desde el punto de vista global, el modelo presenta un valor del log-verosímil de −3015.41 y unos criterios de información AIC = 6034.83 y BIC = 6048.39, coherentes con la solución parsimoniosa seleccionada. Estos valores confirman que el modelo ofrece un compromiso adecuado entre capacidad explicativa y complejidad, en línea con la estrategia de selección basada en BIC.

8.4 Interpretación de los coeficientes

En cuanto a los coeficientes estimados, se observa que la mayoría de las variables incluidas son estadísticamente significativas, lo que indica que aportan información relevante para explicar la probabilidad de que un municipio presente alta abstención electoral. Las variables demográficas y socioeconómicas muestran efectos consistentes con la interpretación sustantiva del fenómeno. Por ejemplo, la variable 'Age_over65_pct' presenta un coeficiente positivo y altamente significativo, lo que implica que un mayor peso de población envejecida se asocia con una mayor probabilidad de abstención elevada. De forma análoga, la tasa de desempleo en el sector de la construcción ('ConstructionUnemploymentPtge') y la variación poblacional ('PobChange_pct') también incrementan significativamente dicha probabilidad.

Las variables territoriales, representadas mediante dummies de comunidad autónoma, capturan efectos regionales no explicados por las covariables estructurales. Destaca el caso de 'CCAA_Cataluña', cuyo coeficiente positivo (2.0948) se traduce en un Odds Ratio de 8.12. Esto indica que, manteniendo constantes el resto de factores, la razón de probabilidad (odds) de registrar niveles elevados de abstención en los municipios catalanes es más de 8 veces superior a la de la categoría de referencia. En contraste, comunidades como Madrid, Castilla-La Mancha o La Rioja presentan coeficientes negativos significativos, lo que sugiere una menor propensión relativa a la abstención alta.

Asimismo, las variables de contexto territorial como 'Densidad_MuyBaja' muestran un efecto protector significativo. Su coeficiente de -0.7059 equivale a un Odds Ratio de 0.494, lo que indica que en los municipios con muy baja densidad poblacional la probabilidad (en términos de odds) de superar el umbral de alta abstención se reduce a menos de la mitad respecto a los de densidad media, una vez controlados el resto de factores.

En conjunto, el modelo final muestra una estructura coherente, con coeficientes estadísticamente significativos. La combinación de variables demográficas, económicas y territoriales permite capturar de manera adecuada los principales determinantes de la abstención elevada, ofreciendo un modelo parsimonioso, estable y con buena capacidad discriminante, adecuado tanto para el análisis explicativo como para la clasificación de municipios en escenarios de riesgo de alta abstención.

```
imp = impVariablesLog(modelo_ganador, y_train, x_train, var_cont_g, var_categ_g,
var_inter_g)

imp
```

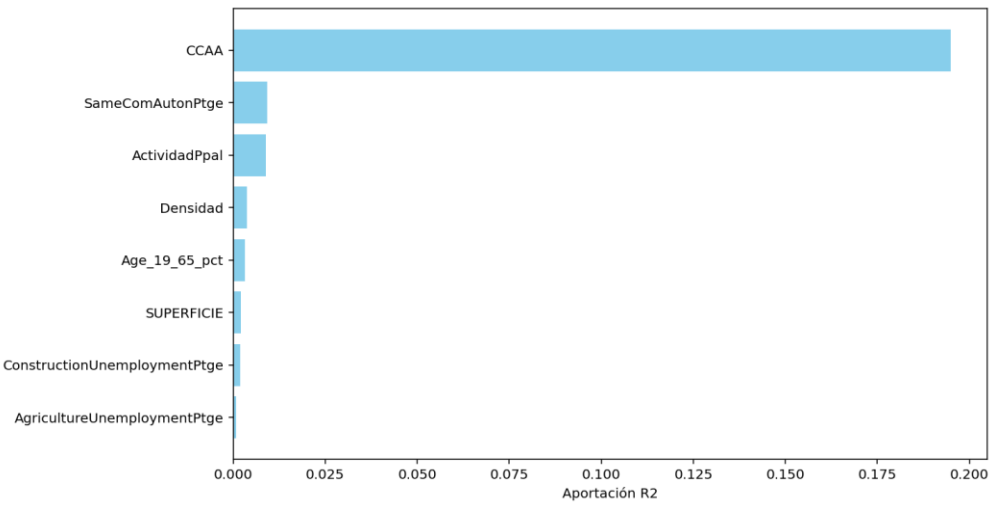

Fig. 11 Importancia de las variables del modelo ganador

La Figura 11 muestra la importancia relativa de las variables incluidas en el modelo final de regresión logística, medida a través de su aportación al pseudo-R^2 del modelo. Esta representación permite evaluar de forma sintética qué grupos de variables contribuyen en mayor medida a la capacidad explicativa global, complementando la interpretación individual de los coeficientes.

Del análisis del gráfico se desprende claramente que el bloque territorial, representado por la variable CCAA, constituye el principal determinante del modelo, con una aportación muy superior al resto de variables explicativas. Este resultado es coherente con los coeficientes estimados en el modelo

logístico, donde varias comunidades autónomas presentan efectos estadísticamente significativos y de gran magnitud. La elevada contribución de 'CCAA' pone de manifiesto la existencia de patrones regionales persistentes en la abstención electoral que no quedan completamente explicados por las variables demográficas o socioeconómicas consideradas.

En un segundo nivel de importancia aparecen variables de carácter estructural y demográfico, como 'SameComAutonPtge', 'ActividadPpal', 'Densidad' y 'Age_over65_pct'. Aunque su aportación individual al pseudo-R^2 es notablemente menor que la del componente territorial, estas variables capturan aspectos relevantes relacionados con la estabilidad residencial, la estructura económica local, el grado de urbanización y el envejecimiento de la población, todos ellos factores tradicionalmente asociados al comportamiento electoral.

Por el contrario, las variables económicas más específicas, como las tasas de desempleo ('ConstructionUnemploymentPtge', 'AgricultureUnemploymentPtge'), así como variables de tamaño ('Superficie', 'PersonasInmueble') o dinámica poblacional ('PobChange_pct'), presentan una contribución marginal al poder explicativo global del modelo. Este resultado no implica que dichas variables carezcan de significación estadística individual, sino que su capacidad adicional para explicar la abstención elevada es limitada una vez controlados los efectos territoriales y demográficos principales.

9. Conclusiones

El análisis desarrollado a lo largo de este trabajo permite concluir que la abstención electoral en España no se distribuye de forma aleatoria, sino que responde a patrones estructurales, territoriales y socioeconómicos bien definidos. La modelización realizada, combinando regresión lineal y regresión logística, ha permitido identificar los principales determinantes del fenómeno y evaluar su capacidad explicativa desde enfoques complementarios.

Desde el punto de vista sustantivo, los resultados ponen de manifiesto una clara jerarquía de factores explicativos.

- El componente territorial emerge como el determinante más relevante: la pertenencia a determinadas Comunidades Autónomas —destacando especialmente Cataluña— incrementa de forma significativa la probabilidad de registrar niveles elevados de abstención. Este resultado sugiere la existencia de factores políticos, institucionales o culturales no observables que influyen de manera diferencial en la movilización electoral y que no quedan plenamente capturados por las covariables estructurales incluidas en el análisis.

- En un segundo nivel de importancia aparecen los factores económicos y demográficos. Las tasas de desempleo, particularmente en los sectores de servicios y construcción, así como el envejecimiento de la población, actúan como catalizadores de la desafección política y se asocian con mayores niveles de abstención. Por el contrario, la baja densidad poblacional muestra un efecto negativo sobre la probabilidad de abstención elevada, lo que sugiere un posible efecto de mayor cohesión social y participación en municipios de menor tamaño.

En términos de modelización, se observa una diferencia clara en el rendimiento de los dos enfoques empleados.

- El modelo de regresión logística ha mostrado un comportamiento robusto y consistente, con valores de AUC superiores a 0.82 y un pseudo-R^2 en torno a 0.25, lo que confirma su utilidad práctica para clasificar municipios con riesgo de alta abstención y apoyar análisis predictivos y de diagnóstico territorial.

- Por su parte, el modelo de regresión lineal, aunque globalmente significativo y con una capacidad explicativa moderada ($R^2 \approx 0.36$), presenta limitaciones asociadas al incumplimiento de algunas hipótesis clásicas, en particular la normalidad de los residuos, lo que aconseja interpretar sus predicciones puntuales con cautela.

Finalmente, conviene señalar algunas limitaciones del estudio. Las restricciones metodológicas impuestas (especialmente la prohibición de transformaciones) han condicionado la capacidad de ajuste de los modelos, sobre todo en el caso de la regresión lineal. Variables de magnitud como población, censo o número de empresas presentan distribuciones fuertemente asimétricas y elevada correlación entre sí, lo que introduce ruido y potencial multicolinealidad. En trabajos futuros, la aplicación de transformaciones logarítmicas, así como la incorporación de términos de interacción (por ejemplo, entre factores económicos y territoriales), permitiría capturar relaciones más complejas y mejorar la capacidad explicativa de los modelos.

A pesar de estas limitaciones, los resultados obtenidos son coherentes, estadísticamente sólidos y sustantivamente interpretables, ofreciendo una aproximación cuantitativa rigurosa al análisis del absentismo electoral municipal y poniendo de relieve el papel central de los factores territoriales en la dinámica de la participación política en España.