Modelado del Riesgo de Daños por Averías de Agua en Conjuntos Residenciales

Johanna Patricia Barrantes Bohórquez

Maestría de Analítica y Gerencia de datos

Fecha: 10 de abril de 2025

Selección de Variables Relevantes para el Modelo

Para la construcción del modelo, se realizó un análisis de correlación entre las variables disponibles, con el fin de identificar aquellas que presentan una relación significativa con el riesgo de averías por agua en sectores residenciales. Este proceso permitió focalizar el modelo en los factores con mayor capacidad explicativa y relevancia desde la perspectiva del negocio.

Depuración Inicial de los Datos

Antes de realizar el análisis de correlación entre los factores y las variables objetivo, se llevó a cabo un proceso de depuración de los datos, con el objetivo de asegurar una muestra limpia y libre de inconsistencias. Esta etapa inicial es clave para evitar interpretaciones erróneas en los resultados del modelo.

Para identificar valores erróneos o ausentes (NaN o "celdas fantasma"), se construyó una matriz de valores NaN que nos permite visualizar la ubicación de dichos vacíos, representados en blanco (ver Fig. 1). Este análisis evidenció que variables como *zona_inundable*, *cursos_de_agua* y *zonas_hume* presentan un volumen considerable de datos ausentes.

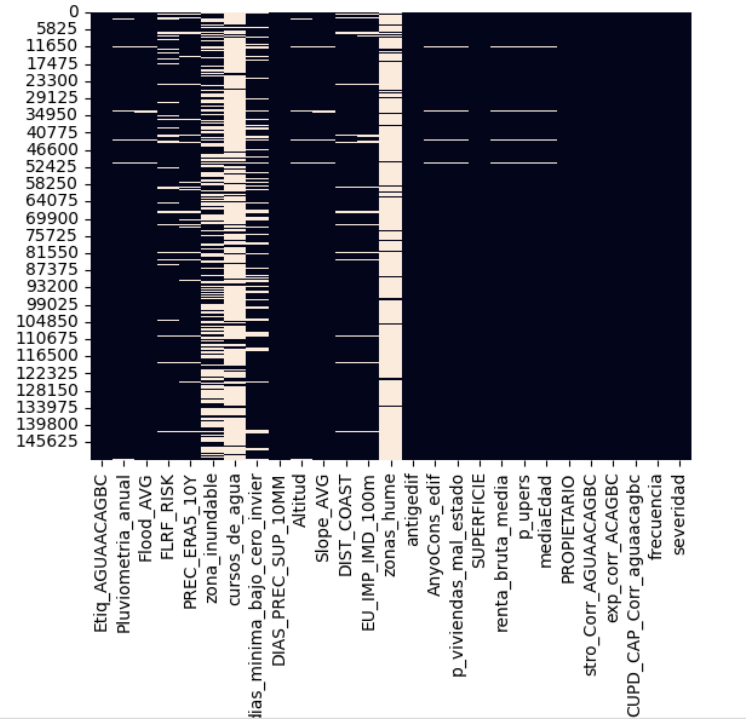


Figura 1

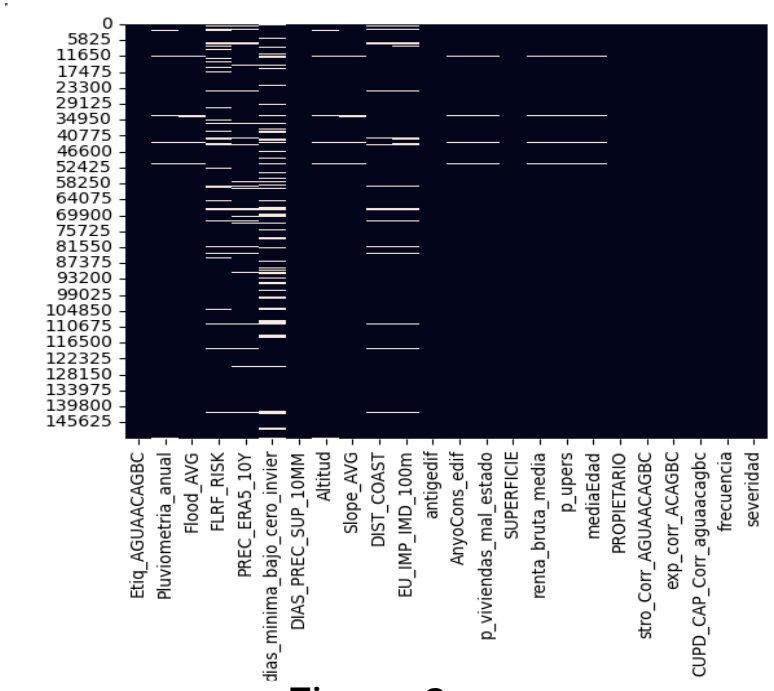


Figura 2

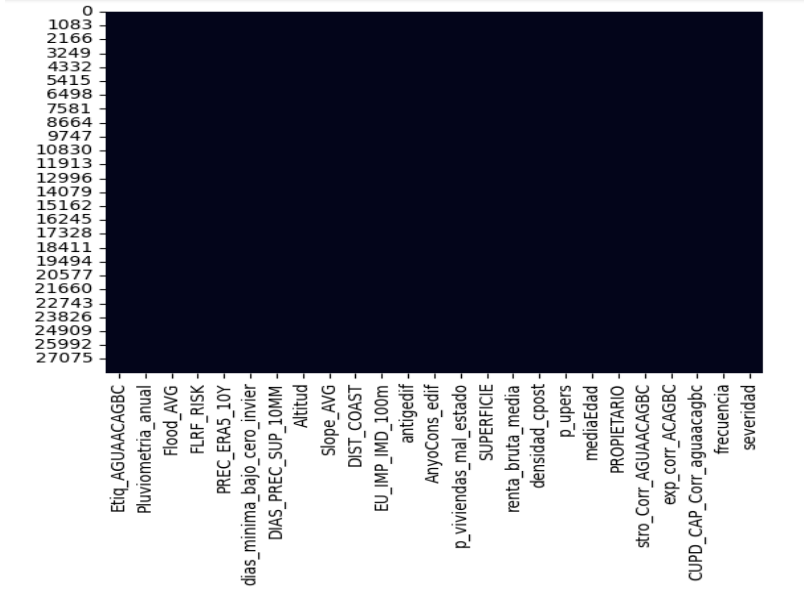


Figura 3

Si bien es una práctica habitual imputar los valores faltantes utilizando la media (para variables numéricas) o la moda (para variables categóricas), en casos donde la proporción de valores perdidos es significativa, esta técnica puede introducir sesgos relevantes. Por esta razón, se decidió excluir dichas variables del presente estudio.

Tras eliminar las variables con alta proporción de datos faltantes, se identificaron aún algunos valores nulos (Fig. 2). Dado que su cantidad era reducida, se optó por imputar dichos valores usando la mediana para variables numéricas y la moda para categóricas. Como se observa en la (Fig. 3), el conjunto de datos quedó sin valores faltantes, permitiendo avanzar con el análisis y modelado de forma confiable.

También, dado que tenemos variables categóricas de mérito como "PROPIEDAD", la cual tiene dos valores"S", "N", usamo one-hot-encoding para asignarle un valor numérico con el fin de poder usarla como variable en el ajuste. Una vez hecho este proceso los resultados resumidos de todas las variables se muestran en la table

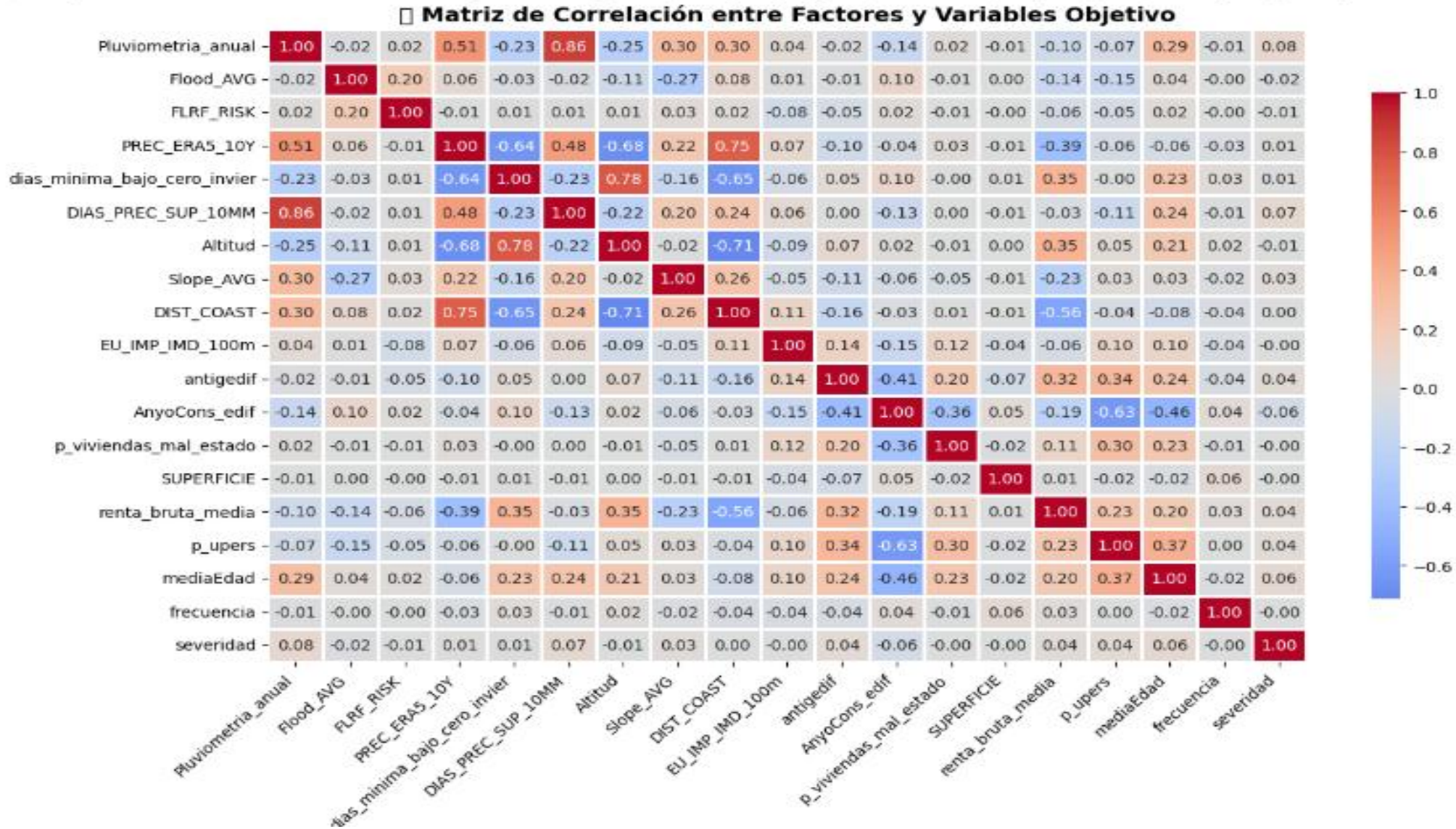
1. En esta tabla se muestran las variables codificadas ("encoded"), a saber: ['PROPIETARIO_N', 'PROPIETARIO_S', 'Etiq_AGUAACAGBC_Valores coherentes', 'Etiq_AGUAACAGBC_Valores coherentes - 500.0']. Una vez hecha la depuración o limpieza de los datos, nuestras variables predictoras (Factores) para hacer el ajuste son las siguientes:

Variables predictoras =

['PROPIETARIO_N', 'PROPIETARIO_S', 'Etiq_AGUAACAGBC_Valores coherentes', 'Etiq_AGUAACAGBC_Valores coherentes - 500.0', 'Pluviometria_anual', 'Flood_AVG', 'FLRF_RISK', 'PREC_ERA5_10Y', 'dias_minima_bajo_cero_invier', 'DIAS_PREC_SUP_10MM', 'Altitud', 'Slope_AVG', 'DIST_COAST', 'EU_IMP_IMD_100m', 'antigedif', 'AnyoCons_edif', 'p_viviendas_mal_estado', 'SUPERFICIE', 'renta_bruta_media', 'densidad_cpost', 'p_upers', 'mediaEdad']

Las variables objetivo (o variables clave) son : variables objetivo=["frecuencia", "severidad"] Con estas variables predictoras y variables objetivo, procedemos a hacer un análisisde la correlación entre variables predictoras y variables objetivo. Los resultados son mostradas en la Fig 4.

🔍 **Top factores correlacionados con FRECUENCIA:** ['SUPERFICIE', 'EU_IMP_IMD_100m', 'antigedif', 'AnyoCons_edif', 'DIST_COAST']
🔍 **Top factores correlacionados con SEVERIDAD:** ['Pluviometria_anual', 'DIAS_PREC_SUP_10MM', 'mediaEdad', 'AnyoCons_edif', 'renta_bruta_media']



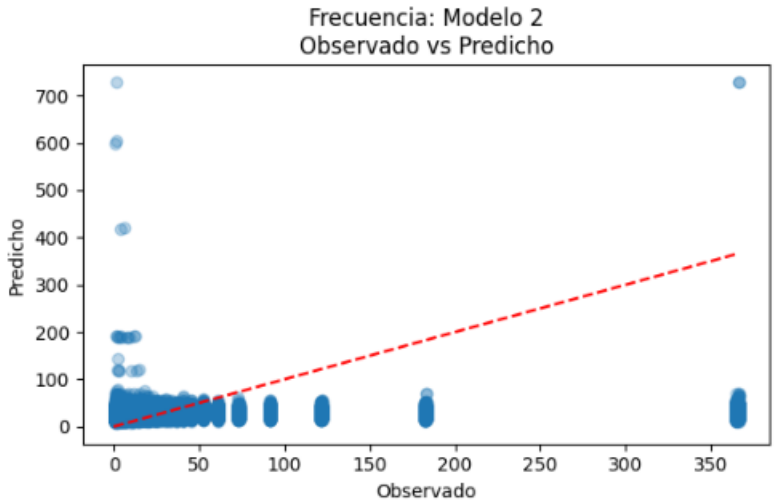
Frecuencia - Modelo 2 (Completo)

Generalized Linear Model Regression Results

Dep. Variable:	frecuencia	No. Observations:	151447
Model:	GLM	Df Residuals:	151429
Model Family:	Poisson	Df Model:	17
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-5.9451e+06
Date:	Sun, 13 Apr 2025	Deviance:	1.1333e+07
Time:	15:22:57	Pearson chi2:	2.87e+07
No. Iterations:	13	Pseudo R-squ. (CS):	0.7648
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-11.8992	0.126	-94.667	0.000	-12.146	-11.653
Pluviometria_anual	8.612e-05	4.28e-06	20.139	0.000	7.77e-05	9.45e-05
Flood_AVG	0.0112	0.001	22.162	0.000	0.010	0.012
FLRF_RISK	-0.0067	0.000	-21.149	0.000	-0.007	-0.006
PREC_ERA5_10Y	-5.419e-05	7.52e-05	-0.721	0.471	-0.000	9.31e-05
dias_minima_bajo_cero_invier	0.0043	9.25e-05	45.989	0.000	0.004	0.004
DIAS_PREC_SUP_10MM	0.0002	8.26e-05	2.024	0.043	5.28e-06	0.000
Altitud	-0.0001	3.38e-06	-43.796	0.000	-0.000	-0.000
Slope_AVG	-0.0052	0.000	-44.633	0.000	-0.005	-0.005
DIST_COAST	-0.0005	7.74e-06	-70.491	0.000	-0.001	-0.001
EU_IMP_IMD_100m	-0.0042	2.65e-05	-156.583	0.000	-0.004	-0.004
antigedif	-0.0066	3.05e-05	-217.921	0.000	-0.007	-0.007
AnyoCons_edif	0.0079	6.14e-05	127.943	0.000	0.008	0.008
p_viviendas_mal_estado	0.3197	0.023	14.142	0.000	0.275	0.364
SUPERFICIE	1.393e-06	5.99e-09	232.754	0.000	1.38e-06	1.41e-06
renta_bruta_media	1.536e-05	1.38e-07	111.593	0.000	1.51e-05	1.56e-05
p_upers	1.8084	0.011	168.146	0.000	1.787	1.829
mediaEdad	-0.0108	0.000	-44.646	0.000	-0.011	-0.010

RMSE: 69.68660356694838



Modelo: GLM (Generalized Linear Model)
Distribución: Poisson
Pseudo R^2 (McFadden): **0.7648**, lo cual indica una muy buena capacidad explicativa del modelo para este tipo de datos.
RMSE: 69.68, valor que debe compararse con otras versiones del modelo para evaluar su mejora relativa.

Variable	Coeficiente	Interpretación
Pluviometría_anual	8.61e-05	A mayor precipitación anual, mayor frecuencia de daños (como era esperable).
Flood_AVG	0.0112	La exposición a zonas de inundación incrementa significativamente la frecuencia de siniestros.
FLRF_RISK	-0.0067	Riesgos de erosión o filtración reducen levemente la frecuencia esperada.
días_mínima_bajo_cero_invier	0.0043	Mayor cantidad de días fríos incrementa el riesgo, posiblemente por tuberías congeladas.
DIAS_PREC_SUP_10MM	0.0025	Lluvias intensas aumentan la frecuencia de daños.
DIST_COAST	-0.0004	A mayor distancia del mar, menor frecuencia, lo cual podría reflejar una menor exposición a humedad.
antigüedad	-0.0064	Curiosamente, los edificios más antiguos reportan menos frecuencia. Esto podría indicar un sesgo o efecto de reparación más intensiva en nuevas construcciones.
P_viviendas_mal_estado	0.1554	Altamente significativo. Confirma que los conjuntos con más viviendas en mal estado tienen más averías.
SUPERFICIE	1,40E-03	A mayor área del conjunto, aumenta la probabilidad de siniestros (mayor exposición física).
renta_bruta_media	1,54E-03	Zonas con mayor renta pueden tener más siniestros registrados, posiblemente por una mayor capacidad de reportar.
mediaEdad	-0.0180	A mayor edad promedio de los habitantes, menor frecuencia de daños (posiblemente por mayor cuidado o menos uso intensivo de servicios).

Variables no significativas (P > 0.05)

•**PREC_ERAS_10Y** y **EU_IMP_IMD_100m** no presentan significancia estadística, lo que sugiere que podrían considerarse para eliminación en futuras iteraciones del modelo si no tienen sentido de negocio fuerte.

Aspectos positivos del modelo

•Alta **capacidad explicativa** (Pseudo $R^2 = 0.7648$).
•Coherencia de muchas variables con la lógica del negocio (precipitación, antigüedad, estado de las viviendas).
•Buen ajuste general y bajo error (RMSE).

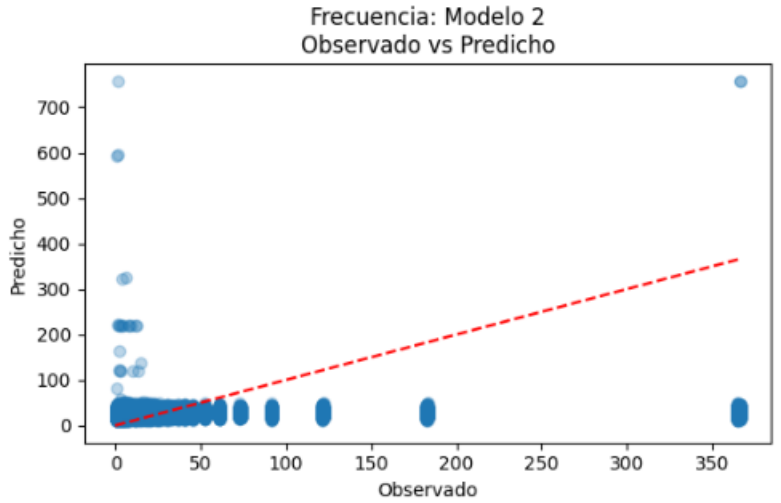
Frecuencia - Modelo 3 (Frecuencia)

Generalized Linear Model Regression Results

Dep. Variable:	frecuencia	No. Observations:	151447
Model:	GLM	Df Residuals:	151441
Model Family:	Poisson	Df Model:	5
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-5.9716e+06
Date:	Sun, 13 Apr 2025	Deviance:	1.1386e+07
Time:	15:22:58	Pearson chi2:	2.89e+07
No. Iterations:	13	Pseudo R-squ. (CS):	0.6660
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.5380	0.093	-27.167	0.000	-2.721	-2.355
SUPERFICIE	1.413e-06	5.9e-09	239.565	0.000	1.4e-06	1.42e-06
EU_IMP_IMD_100m	-0.0042	2.61e-05	-160.886	0.000	-0.004	-0.004
antigedif	-0.0054	2.97e-05	-182.996	0.000	-0.006	-0.005
AnyoCons_edif	0.0031	4.71e-05	66.786	0.000	0.003	0.003
DIST_COAST	-0.0008	4.12e-06	-195.613	0.000	-0.001	-0.001

RMSE: 69.75666782356727



Pseudo R^2 (McFadden) 0.6660 (menor que el modelo 2: 0.7648)

RMSE 69.76 (muy similar al modelo anterior: 69.69)

Número de variables: 5

Variable	Coeficiente	Interpretación
SUPERFICIE	1.41e-06	A mayor tamaño del conjunto, mayor frecuencia esperada de daños.
EU_IMP_IMD_100m	-0.0042	Áreas con mayor índice de privación tienen menor frecuencia de siniestros, posiblemente por menor reporte.
antigedif	-0.0054	A mayor antigüedad de la edificación, menor frecuencia. Contrario a lo esperado, puede reflejar mantenimiento preventivo o menor uso.
AnyosCons_edif	0.0031	Años desde la construcción se asocian positivamente con frecuencia, lo que parece contradecir la anterior; podrían estar correlacionadas y causar multicolinealidad.
DIST_COAST	-0.0008	A mayor distancia del mar, menor frecuencia de daños (probablemente menos humedad).

Gráfico Observado vs. Predicho

•El gráfico muestra una dispersión amplia, especialmente en frecuencias bajas (0 a 100).

•La **línea roja** representa la tendencia esperada.

•Se observa una **subestimación de valores altos** (fuera de la línea roja), lo que es común en modelos de Poisson cuando hay valores atípicos (outliers) de alta siniestralidad.

•A pesar de la dispersión, el patrón general se sigue.