Modelo GLM para averías relacionadas con agua en conjuntos residenciales.

Presentado por: Paula Alejandra Acero Hoyos

MAESTRÍA EN ANALÍTICA Y GERENCIA DE DATOS

P **por Paula Acero**

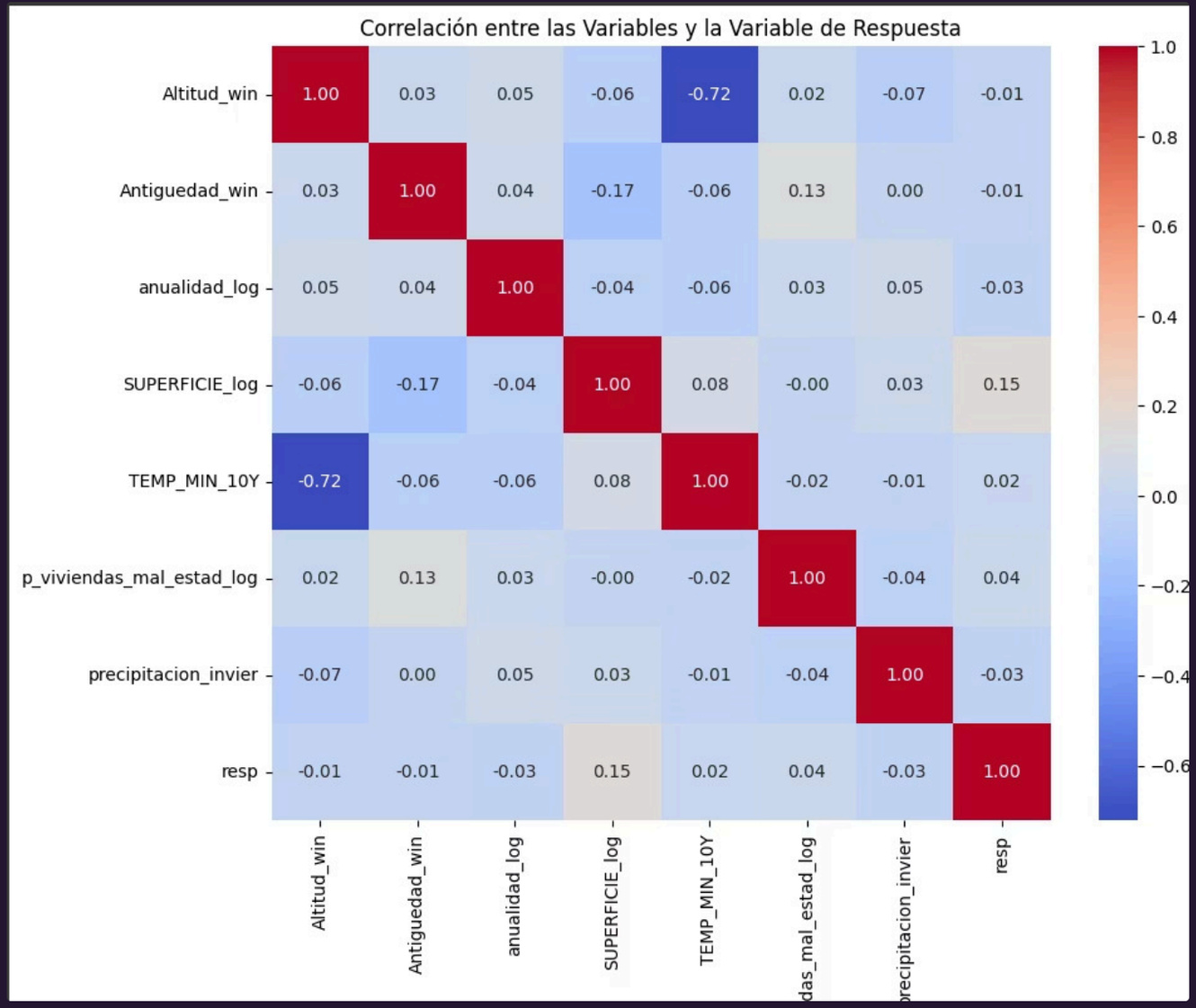
Objetivo

El objetivo de este modelo es **predecir el número de siniestros relacionados con averías por agua en conjuntos residenciales,** en función de varias variables explicativas (como la **antigüedad de las viviendas, la superficie construida**, etc.) y entender qué factores afectan significativamente la frecuencia de estos siniestros.

Limpieza de variables

- Se reemplazaron los valores nulos y erróneos (como -9999 o 0) por valores de reemplazo como NaN, mediana, o logaritmos.
- Se aplicó **winsorización** para controlar los outliers.
- Se realizaron **transformaciones logarítmicas** en variables sesgadas para estabilizar la varianza y hacer las relaciones más lineales.

Correlación entre las variables



MODELO GLM PARA LA FRECUENCIA

En este caso, fue escogio el modelo 2, ya que presenta un BIC más bajo.

Generalized Linear Model Regression Results

Dep. Variable:	resp	No. Observations:	51319
Model:	GLM	Df Residuals:	51312
Model Family:	Poisson	Df Model:	6
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1.5230e+06
Date:	Sun, 13 Apr 2025	Deviance:	3.0074e+06
Time:	15:42:07	Pearson chi2:	1.66e+09
No. Iterations:	10	Pseudo R-squ. (CS):	0.9997
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.5470	0.019	-348.948	0.000	-6.584	-6.510
Altitud_win	-0.0001	7.06e-06	-19.027	0.000	-0.000	-0.000
Antiguedad_win	0.0064	0.000	52.024	0.000	0.006	0.007
anualidad_log	-0.3337	0.004	-94.608	0.000	-0.341	-0.327
SUPERFICIE_log	1.2440	0.002	639.590	0.000	1.240	1.248
p_viviendas_mal_estad_log	9.2280	0.071	130.496	0.000	9.089	9.367
precipitacion_invier	-0.1478	0.001	-99.971	0.000	-0.151	-0.145

Análisis del resultado del modelo:

Las variables como p_viviendas_mal_estado_log (porcentaje de viviendas en mal estado) y SUPERFICIE_log (superficie construida) muestran una relación fuerte con la variable respuesta.

- p_viviendas_mal_estado_log : A medida que aumenta el **porcentaje de viviendas en mal estado**, es más probable que haya **más siniestros**, ya que las viviendas en mal estado podrían tener más problemas estructurales, aumentando el riesgo de siniestros.
- SUPERFICIE_log : A mayor **superficie construida**, el riesgo de siniestros aumenta, ya que áreas más grandes tienen más instalaciones y, por lo tanto, más probabilidad de sufrir daños.

Análisis del resultado del modelo:

- En esta sección, se destacan las variables que tienen un **efecto estadísticamente significativo** en el número de siniestros, según los coeficientes del modelo.
- **Coeficientes positivos** y **negativos** indican **cómo** y **cuánto** las variables afectan la cantidad de siniestros.

- Antiguedad_win : Cada aumento en la antigüedad de la vivienda está asociado con un **aumento en los siniestros**, lo que tiene sentido porque las viviendas más antiguas pueden ser más propensas a fallos estructurales.
- anuelidad_log : Un mayor valor de la **anualidad** está relacionado con **menos siniestros**, lo que podría indicar que los seguros con mayor cobertura o precios son más completos y cubren mejor los riesgos.

Análisis del resultado del modelo:

- Altitud_win : A medida que aumenta la altitud, disminuye la frecuencia de los siniestros. Esto tiene sentido si consideramos que las áreas más altas pueden tener menos riesgo debido a otros factores.
- precipitacion_inv : El coeficiente negativo sugiere que la precipitación en invierno disminuye la frecuencia de los siniestros.