Proyecto Final Pricing
Juan David Sierra

1. Objetivo del Proyecto

Desarrollar un modelo estadístico interpretable que permita predecir la frecuencia de siniestros por daños de agua en conjuntos residenciales, utilizando información climática, estructural, sociodemográfica y de siniestralidad histórica.

2. Flujo de Desarrollo del Modelo

Carga y limpieza de datos

- Se consolidaron múltiples archivos .parquet.
- Se reemplazaron valores codificados como -999 por NaN.
- Se eliminaron columnas con más del 30% de datos faltantes y aquellas sin variabilidad.
- Se codificaron variables categóricas numéricamente para facilitar el modelado.

Definición de la variable objetivo
Se creó la variable de respuesta como:

```
df = df[df['exp_corr_ACAGBC']>0]
df['frecuencia'] = df['stro_Corr_AGUAACAGBC']/df['exp_corr_ACAGBC']
```

Solo se conservaron registros donde el denominador era mayor que cero.

3. Selección de Variables

Rol de Random Forest
Se utilizó un modelo de Random Forest Regressor con fines exploratorios para identificar las variables más relevantes sin necesidad de conocimiento de negocio previo.

- Se identificaron 3 variables con mayor importancia: EXPOSICION, CUPD_CAP_Corr_aguaacagbc y K_ACAGBC

- Se excluyeron variables relacionadas directamente con la fórmula de frecuencia para evitar fuga de información.

Análisis de colinealidad
Se aplicó la métrica **VIF (Variance Inflation Factor)** sobre las variables más importantes:

- Variables con VIF > 10 fueron eliminadas para evitar colinealidad entre predictores.
- El modelo final incluyó solo aquellas con VIF ≤ 10.

4. Modelado Estadístico: GLM (Poisson)

Se entrenó un Modelo Lineal Generalizado (GLM) con distribución Poisson sobre las variables filtradas.

Métricas de desempeño:

- **AIC:** 597,399.23 → indicador de calidad del modelo penalizado por complejidad.
- **Deviance:** 566,123.04 → bondad de ajuste general del modelo.
- **Pseudo R^2 (CS):** ~1.0 → fuerte capacidad explicativa sobre los datos.

Interpretación de variables principales:

Variable	Coef.	Interpretación
EXPOSICION	-13.20	Mayor exposición reduce la frecuencia (mantenimiento eficaz)
CUPD_CAP_Corr_aguaacagbc	+0.0008	Mayor costo → más siniestros reportados
K_ACAGBC	+8.6e-07	Riesgo de agua → positivo y consistente
Contr_IndefH_pct	-0.056	Contratos indefinidos → menos siniestros
porcentaje_no_residente	+0.0020	Más no residentes → más frecuencia por menor cuidado
Elevation_AVG	+0.0003	Altura → ligera correlación con más siniestros
p_suelo_uso_comercial	+1.97	Uso comercial → mayor tráfico o deterioro estructural

El modelo GLM con distribución Poisson logró identificar 8 variables predictoras sin multicolinealidad, todas estadísticamente significativas, que explican la frecuencia de siniestros por agua. Se utilizó un enfoque automatizado con Random Forest y VIF para seleccionar variables objetivamente. El modelo mostró un buen ajuste (AIC bajo y pseudo $R^2 \approx 1$), con patrones lógicos y defendibles como mayor frecuencia en zonas con mayor exposición, costos de reparación altos y menor cuidado habitacional.

5. **Visualizaciones Generadas**

- Gráfico de importancia de variables (Random Forest)
- Gráfico de VIF para detectar colinealidad
- Gráfico de valores observados vs. predichos (GLM) para evaluar el ajuste