

Actividad Final: Análisis y Modelación del Seguro de Automóviles

Luis Alejandro Ferro Alfonso

Electiva I

Contexto General

La siguiente actividad utiliza una base de datos suministrada por una **compañía aseguradora**, compuesta originalmente por cientos de miles de pólizas reales de seguro de automóviles.

Con el fin de proteger la confidencialidad de los asegurados y cumplir con los lineamientos de manejo seguro de información, la aseguradora generó una **base de datos simulada de 389.654 pólizas**, preservando únicamente los patrones estadísticos, distribuciones y relaciones estructurales observadas en los datos reales.

La base final se entrega dividida en cuatro archivos debido a su tamaño.

Los datos reflejan:

- Estructura real del parque automotor asegurado.
- Patrón actuarial del comportamiento de los conductores.
- Tendencias reales de siniestralidad según marca, línea, color, uso y perfil del conductor.

Descripción de Variables

La base contiene las siguientes variables:

1. Información del vehículo

- **id_poliza**: Identificador único de 10 dígitos.
- **clase_veh**: Automóvil o camioneta.
- **marca**: Marca del vehículo (Renault, Chevrolet, Mazda, etc.).
- **linea**: Línea o referencia exacta.

- **modelo:** Año de fabricación.
- **cilindraje:** Cilindraje del motor.
- **antiguedad:** Años desde el modelo hasta 2025.
- **valor_asegurado:** Valor comercial aproximado basado en tablas públicas.
- **uso_vehiculo:** Particular, empresa, taxi, plataforma, carga.
- **km_anual:** Kilometraje anual estimado según el uso.
- **dispositivos_seguridad:** Ninguno, alarma, GPS o ambos.
- **color_vehiculo:** Color del vehículo.

2. Información del conductor

- **edad_conductor**
- **genero**
- **estado_civil**
- **anios_experiencia**
- **hist_siniestros_3a:** Siniestros en últimos 3 años.
- **num_multas_2a:** Multas en últimos 2 años.
- **dias_uso_semana:** Frecuencia semanal de utilización.

3. Componentes actuariales simulados

- **exposicion:** Porcentaje del año asegurado.
- **tipo_cobertura:** Terceros o Todo Riesgo.
- **num_siniestros_anio:** Frecuencia anual simulada.
- **costo_total_siniestros:** Severidad agregada.

Parte I: Análisis Descriptivo y Descubrimiento de Patrones

El objetivo es que los estudiantes identifiquen patrones en siniestralidad y comportamiento del mercado.

Actividades

1. Realice una descripción estadística inicial de las variables cuantitativas.
2. Construya gráficos comparativos por:
 - marca,
 - línea,
 - color,
 - uso del vehículo,
 - edad del conductor.
3. Identifique y explique:
 - qué marcas presentan mayor siniestralidad,
 - qué colores tienen mayor frecuencia,
 - qué perfiles de conductor presentan mayor riesgo,
 - si el uso del vehículo modifica el comportamiento del riesgo.
4. Señale al menos **5 patrones ocultos** que usted descubra en la base.

Parte II: Modelación Actuarial del Seguro

El objetivo es que los estudiantes definan su propio modelo actuarial para calcular la prima pura de riesgo usando la base de datos.

1. Frecuencia

Sea N_i el número de siniestros del vehículo i .

Cada estudiante debe estimar:

$$\hat{\lambda}_i = E[N_i \mid X_i]$$

Posibles enfoques:

- Poisson GLM
- Binomial Negativa
- Tarificación por clases (rating factors)
- Tarificación por relatividades

Las variables sugeridas para frecuencia:

$$X_i = (\text{edad}, \text{marca}, \text{uso}, \text{color}, \text{historial}, \text{multas})$$

2. Severidad

Sea Y_{ij} el costo del siniestro j del vehículo i .

Se solicita modelar:

$$E[Y_{ij} \mid X_i]$$

Enfoques posibles:

- Lognormal GLM
- Gamma GLM
- Ajustes paramétricos con MLE

Variables sugeridas:

$$X_i = (\text{valor asegurado, clase, edad, tipo de cobertura})$$

3. Prima Pura de Riesgo

Cada estudiante debe calcular:

$$\text{Prima Pura}_i = \hat{\lambda}_i \times \hat{E}[Y_i]$$

donde:

$$\hat{E}[Y_i] = \text{Severidad Esperada}$$

4. Prima de Tarifa

Si desea, puede agregar:

$$\text{Prima Tarifa}_i = \text{Prima Pura}_i \times (1 + \text{Gastos} + \text{Utilidad})$$

Sugiera un recargo razonable.

5. Entregable final

Cada estudiante debe entregar un **informe en formato PDF**, generado exclusivamente desde **RMarkdown**. El documento debe ser completamente reproducible e incluir tanto el código en **R** como los resultados obtenidos. El informe debe contener:

- Estimación de la frecuencia esperada por póliza (λ_i).
- Estimación de la severidad esperada.
- Cálculo de la prima pura individual y su justificación actuarial.
- Gráficos de sensibilidad y análisis exploratorio.

- Discusión técnica: identificación de los factores que más contribuyen al riesgo.

Además del informe, cada grupo de estudiantes deberá realizar una **presentación corta (máximo 10 minutos)** donde exponga de manera clara y visual:

- Los hallazgos principales de la **Parte 1: análisis exploratorio y descubrimiento de patrones**.
- Los resultados de la **Parte 2: construcción de modelos de frecuencia, severidad y tarificación**.
- Los aspectos más relevantes, sorpresas, patrones ocultos encontrados y conclusiones a nivel actuarial.

La presentación se elabora en R Markdown y puede elaborarse en dos formatos:

1. **Beamer (LaTeX)**, o
2. **Presentación HTML generada desde RMarkdown** (`html_document` o `xaringan`).

Se valorará la claridad, la capacidad de síntesis, el uso adecuado de visualizaciones y la interpretación técnica de los resultados.