**Proyecto Asignatura Ciencia de Datos**

**Entregable #2:** Inventario de Datos

Cristian Enrique Munar Mora

**Planear el Inventario:**

* **Propósito del Inventario:** El propósito de este inventario de datos es considerar, enunciar y describir todas las tablas (*data-sets*) de que se dispone para la resolución del problema planteado: *¿Cómo puede la compañía mejorar su resultado del Ramo de SOAT a partir de una mejor selección de riesgos?* […] Problema que se reduce al entendimiento de *¿Qué variables influyen en la distribución de Frecuencia y Costo Medio de los siniestros del Ramo de SOAT?*

Se busca también con este inventario entender qué dificultades puede tenerse para el acceso a la data que se necesita y las características técnicas de la misma: almacenamiento, frecuencia de actualización, dueños y gestores de la data, etc.

* **Sostenibilidad del Inventario:** Las diferentes fuentes de datos, como se describirá en el Anexo 1 de este entregable, tienen su origen en la operación del negocio y en la transformación de los datos capturados en los procesos de cotizaciones, emisiones, siniestros, y otros, en información útil para el seguimiento del resultado técnico y financiero del ramo de SOAT. En este sentido, la información utilizada para este proyecto está siendo actualizada en periodos no mayores a un mes en el caso menos frecuente.
* **¿Qué Incluir?** La base de la cual se partirá para el poblamiento de la data requerida de este proyecto corresponde al universo de todas las pólizas que estuvieron vigentes en la compañía para el ramo de SOAT (en determinada ventana de tiempo), así como la siniestralidad asociada a estas pólizas.

Es preciso pues plantear ciertas definiciones necesarias para el entendimiento del problema:

*Amparo: Cada uno de los riesgos o eventos que están cubiertos por una póliza de seguros. En el caso de una póliza de SOAT los amparos son: Gastos Médicos, Incapacidad Permanente, Gastos de Transporte, Muerte y Auxilios Funerarios.*

*Expuesto: Porción del riesgo que se encuentra vigente en determinada unidad de tiempo. Por definición 1 expuesto corresponde a la vigencia completa de una póliza (un año). Así, por ejemplo, los expuestos calculados para una póliza en un mes, serán igual a 1/12.*

*Casos: Cantidad de siniestros ocurridos sobre una cartera de pólizas en un periodo determinado. En el ramo de SOAT un único evento (siniestro/accidente) puede dar lugar a varias víctimas. Ejemplo: Accidentes múltiples o choques en los que haya más de un muerto o lesionado.*

*Valor Incurrido: Corresponde al monto pagado o reservado para el cumplimiento de las obligaciones derivadas de los casos ocurridos en un periodo determinado. Es necesario mencionar que el producto de SOAT está diseñado para asumir hasta cierto monto máximo por victima en cada uno de los amparos: Ejemplo: 800 salarios mínimos diarios legales vigentes (SMDLV) para la cobertura de Gastos Médicos.*

*Frecuencia: Relación existente entre la cantidad de casos ocurridos sobre la cantidad de expuestos en un periodo determinado.*

*Costo Medio: Valor que, en promedio, le cuesta a la compañía cada una de las víctimas para un ramo o amparo determinado.*

*Prima Pura de Riesgo: Valor esperado del monto de la reclamación para cualquier póliza del Ramo. Resulta del producto de la Frecuencia (proporción de pólizas que se siniestran) y el Costo Medio (monto promedio pagado para cada póliza dado que esta se ha siniestrado).*

Adicionalmente a esta información de carácter técnico, se requerirá recopilar todas aquellas variables a las que se puede acceder en las bases de datos de la compañía y de las cuales se considere puedan ser predictivas para la siniestralidad: Características del Asegurado, Características del Vehículo, Zona de Circulación, Canal de Venta, etc.

* Nivel de Detalle Requerido: Según experiencias previas y conocimiento del negocio, para abordar este problema, se considera apropiado tomar la experiencia siniestral de los últimos 3 años. No obstante, debido a las medidas preventivas y restrictivas establecidas por el Gobierno Nacional para minimizar los efectos del COVID en Colombia, el 2020 fue un año atípico en términos de siniestralidad. En este sentido, resulta conveniente omitir la información de este último año por el sesgo que pueda generar en los resultados obtenidos.

Teniendo en cuenta lo anterior, se tomará toda la información de expuestos, casos e incurrido para los años 2017, 2018 y 2019 omitiendo así los efectos derivados del COVID.

**Descripción del Inventario de Datos:**

Ver Anexo 1: Entregable #2 – Data Inventory.xlsx