Sura Digital CLV: Valor del Cliente con Inteligencia de Datos

# Parte 1

# Comprensión del Negocio

En el contexto actual del mercado asegurador, especialmente en el ámbito digital de los seguros voluntarios enfocados en automóviles, es fundamental adoptar una visión estratégica que trascienda la mera adquisición de nuevos clientes. Reconocemos que el verdadero valor para la organización no reside únicamente en incrementar el número de usuarios, sino en identificar, conservar y fidelizar a aquellos clientes que, por sus características y comportamiento, representan un alto valor potencial para el negocio a largo plazo.

**Objetivos específicos**

• Obtener y consolidar datos relevantes del ecosistema digital de Sura asociados al comportamiento de los clientes del segmento de movilidad, incluyendo transacciones, interacciones digitales, historial de compras, renovaciones, uso de servicios y características demográficas.

• Preprocesar, limpiar y transformar los datos recolectados, manejando valores nulos y outliers, así como codificación de atributos categóricos relevantes para el modelo.

• Calcular el Customer Lifetime Value (CLV) histórico de los clientes del canal digital de Sura correspondientes a los últimos dos años, utilizando los datos consolidados de primas devengadas, siniestros incurridos y costos de asistencias, con el fin de establecer una métrica base confiable para análisis predictivo y segmentación.

• Construir modelos predictivos supervisados para estimar el CLV futuro de los clientes, utilizando algoritmos como regresión lineal, árboles de decisión, random forest, XGBoost o redes neuronales, evaluando su desempeño con métricas como MAE, RMSE y R².

• Entregar al gerente del canal digital de Sura un informe ejecutivo con los resultados del modelo de CLV, con el fin de respaldar la toma de decisiones estratégicas relacionadas con la inversión en pauta digital y las acciones de fidelización de clientes.

# Comprensión de los Datos

El data set *Data\_Analisis\_Movilidad\_*U2Y.csv para el desarrollo del proyecto fue obtenido desde la compañía Suramericana de Seguros, donde se valida información de los clientes del canal digital de los últimos 2 años, con el fin de proteger a información sensible de los clientes es fundamental considerar desde la fase de diseño los lineamientos establecidos por la Ley de Habeas Data. Por esto se anonimizan datos de algunas columnas (Dni Aseg Anonymized, Asegurado Id Anonymized, Poliza Certificado Id Anonymized, Numero Poliza Anonymized, Placa Vehiculo Anonymized y Nombre Agente Anonymized).

Esta información reposa en la nube de la compañía, donde se descarga por medio de un query en *Teradata* solicitando los datos de los clientes del canal digital para posteriormente ser tratados.

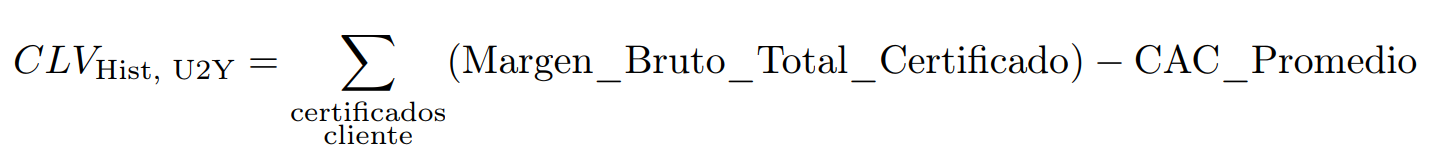
El *dataset* tiene 45 columnas y 22.904 registros, los cuales contienen información demográfica de clientes (edad, sexo, estado civil, ubicación), detalles de pólizas de seguro, información del vehículo asegurado, y métricas financieras relacionadas con primas y siniestros.



Nuestra Variable predictora la mediremos mediante el calculo de (CLV)

Dado que la tabla *Data\_Analisis\_Movilidad\_U2Y.csv* contiene métricas financieras (Prima\_Total\_Devengada\_Certificado,Siniestros\_Total\_Incurridos\_Certificado, Costo\_Total\_Asistencias\_Certificado) que cubren la vida completa de cada certificado iniciado en los últimos dos años, un cálculo adecuado del CLV histórico

a nivel de cliente, basado en estos datos, sería:



Donde:

**Σcertificados cliente:** Indica la suma de los valores para todos los certificados (filas en Data\_Analisis\_Movilidad\_U2Y.csv) pertenecientes al mismo cliente (Dni\_Aseg\_Anonymized).

**Margen\_Bruto\_Total\_Certificado:** Es el campo calculado en la tabla final, que representa (Prima Total - Siniestros Totales - Costos Asistencias Totales) para la vida de ese certificado específico.

**CAC\_Promedio:** Es el Costo de Adquisición de Cliente promedio (el cual está definido de acuerdo a los costos asociados al costo por lead y costo en pauta Vs numero de pólizas expedidas). S e resta una sola vez por cliente.

**Lógica y justificación**

• Agrega el valor bruto generado por todas las pólizas/certificados que el cliente inició en el canal Directo Digital durante los últimos dos años, considerando la rentabilidad de cada una durante su vida completa (hasta la fecha o su finalización).

• Restar el CAC promedio proporciona una estimación del valor neto histórico generado por las interacciones recientes (últimos 2 años) del cliente en este canal específico.

• Es un enfoque histórico basado en datos pasados.

• Sirve como base para el modelo predictivo, pudiendo ser la variable objetivo (target) a predecir.

• Limitación Clave: Refleja el valor de las pólizas iniciadas en los últimos 2 años. La contribución histórica de pólizas anteriores del mismo cliente en el canal no está incluida aquí.

# 

El *dataset* está completo, sin valores nulos, pero presenta una anomalía en la variable Edad\_Cliente. La mayoría de las variables muestran distribuciones sesgadas, especialmente Num\_Total\_Asistencias\_Certificado, que tiene una media baja, pero valores extremos altos, indicando clientes con muchas asistencias puntuales. Las variables de tipo binario evidencian una baja adopción de productos como Soat, Hogar o Salud, lo que puede ser clave para segmentación.

# Preparación de los Datos

**Limpieza**

En esta etapa se aplicaron acciones clave de limpieza y transformación de datos con el objetivo de asegurar la calidad, coherencia y relevancia del conjunto de datos para análisis posteriores. Las operaciones realizadas fueron:

* Originalmente contenían valores numéricos representados como texto con formato regional, es decir, usando puntos como separadores de miles y comas como separadores decimales. Para estandarizar estos datos al formato numérico requerido por Python (donde el punto es el separador decimal).  
    
  Este procedimiento asegura que variables como *Valor\_Comercial\_Vehiculo, Margen\_Tecnico\_Total\_Certificado* o *Ultima\_Prima\_Anualizada\_Cert*, entre otras, puedan ser tratadas correctamente como valores cuantitativos, eliminando inconsistencias derivadas del formato de entrada y habilitando análisis posteriores precisos y confiables.
* En las columnas *Valor\_Accesorios\_Vehiculo\_Sia* y *Valor\_Accesorio\_Especial\_Veh*

contenían valores no numéricos representados por el carácter "?", se sustituyeron todos por 0, asumiendo que la ausencia de un valor representa la inexistencia de un accesorio o valor monetario asociado.

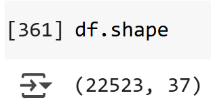
**Tratamiento de la data**

* En la columna ‘Edad\_Cliente’ presenta 415 registros con valores ‘-1’ ó inferiores a 18 años, por la cantidad de registros, los cuales son alrededor del 1.8% del total de la data se toma la decisión de reemplazarlos por la media de la columna.
* La Columna ‘segmento\_edad’ presenta 417 registros con valores ‘?’ los cuales se reemplazan por la moda de la columna.
* En la columna ‘Sexo\_Cd’ se encontraron 382 registros con valores ‘-1’ los cuales fueron eliminados ya que no tenían mayor representación en el conjunto de datos.
* En las columnas Ciudad\_Cliente, Departamento\_Cliente, Sector\_Economico\_Cliente, se eliminan los registros ‘Sin\_Información’, estos son registros que representan menos del 1% de la data.
* Se excluyeron planes específicos de la variable ‘*Subproducto\_Desc’* que no aplicaban al análisis o representaban menos del 1% del total de registros, como 'PLAN BICISURA', 'No Aplica' y 'Plan Seguro Muévete Libre', con el fin de evitar sesgos en el modelo.
* Se eliminan algunas variables que no serán utilizadas en el dataset, los cuales son los valores anonimizados y la columna ‘Rango\_Ingresos\_Cliente’ y ‘Estado\_Civil\_Desc’, estas últimas ya que había 18.744 registros sin información.
* En las columnas ‘Valor\_Accesorios\_Vehiculo\_Sia’ y ‘Valor\_Accesorio\_Especial\_Veh’ se reemplazan los registros con valores ‘?’ y se reemplazan por ‘0’

**Conversión de variables binarias numéricas a categóricas legibles:**

* Las variables relacionadas con productos activos del cliente *(Cliente\_Tiene\_Soat\_Activo', 'Cliente\_Tiene\_Hogar\_Activo', 'Cliente\_Tiene\_MasVida\_Activo','Cliente\_Tiene\_Salud\_Activo','Cliente\_Tiene\_Vida\_Activo')* fueron transformadas de valores numéricos (0 y 1) a etiquetas categóricas **('No' y 'Sí').** Esto mejora la interpretabilidad y utilidad en visualizaciones y modelos de clasificación.

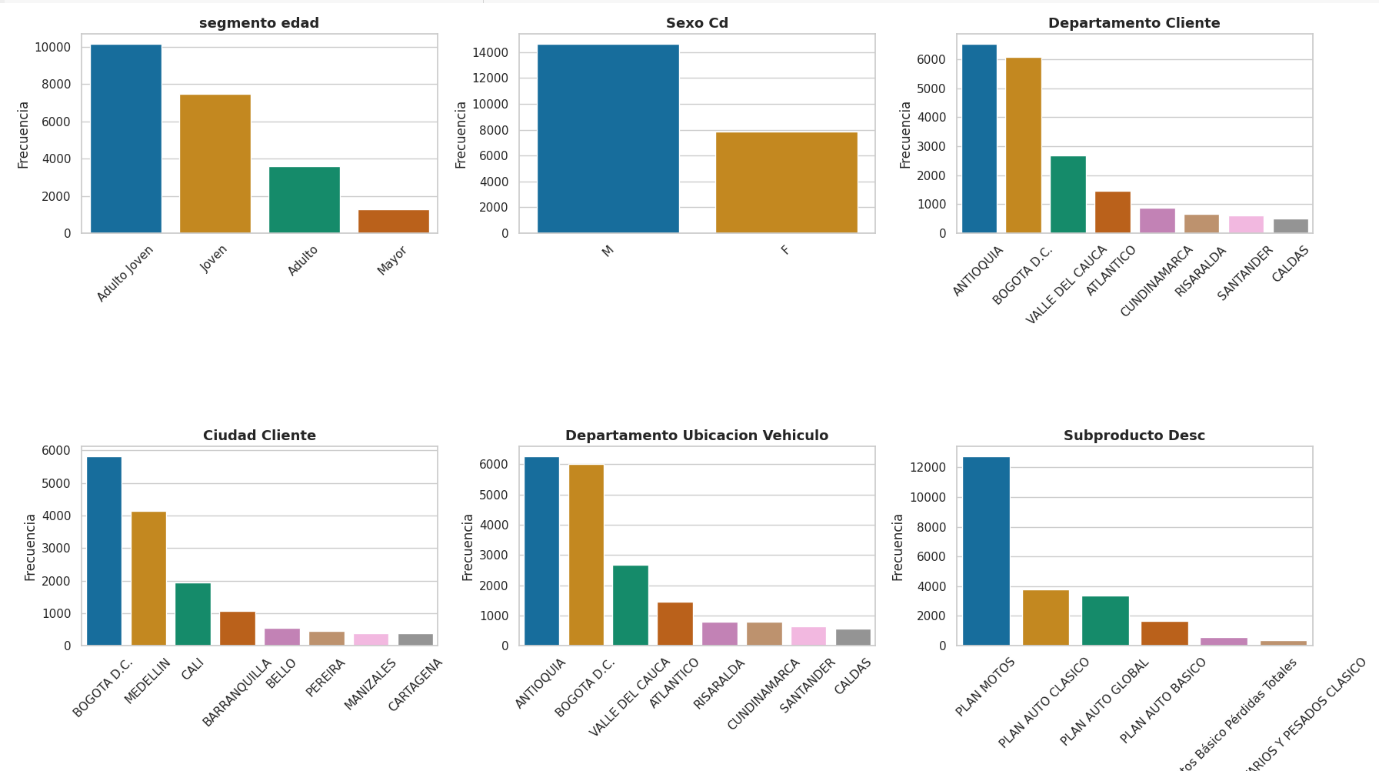
Este conjunto de transformaciones permitió refinar el data set, eliminando ruido y garantizando que los datos restantes sean representativos, consistentes y aptos para análisis exploratorios, segmentaciones o modelado predictivo.

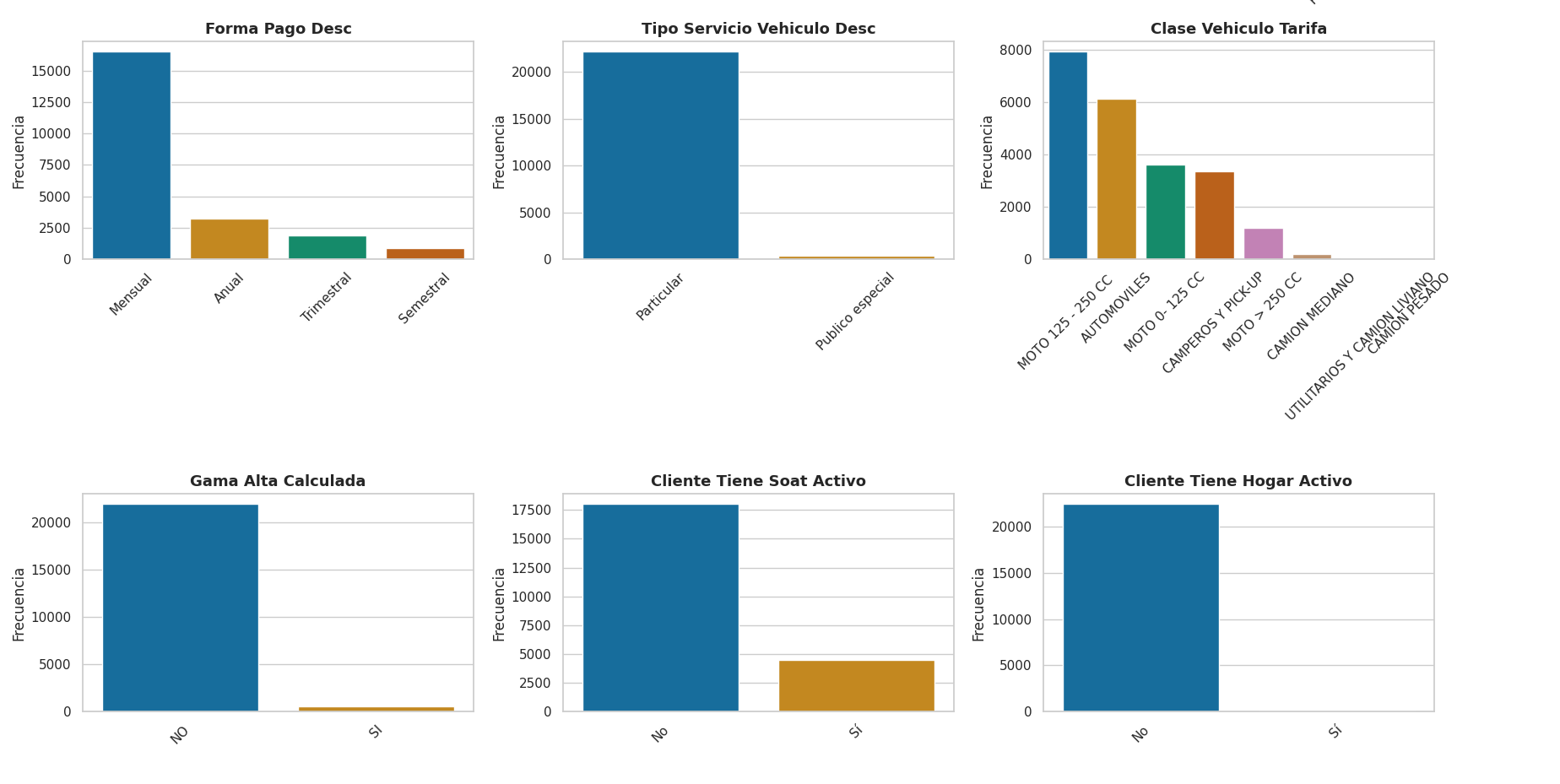


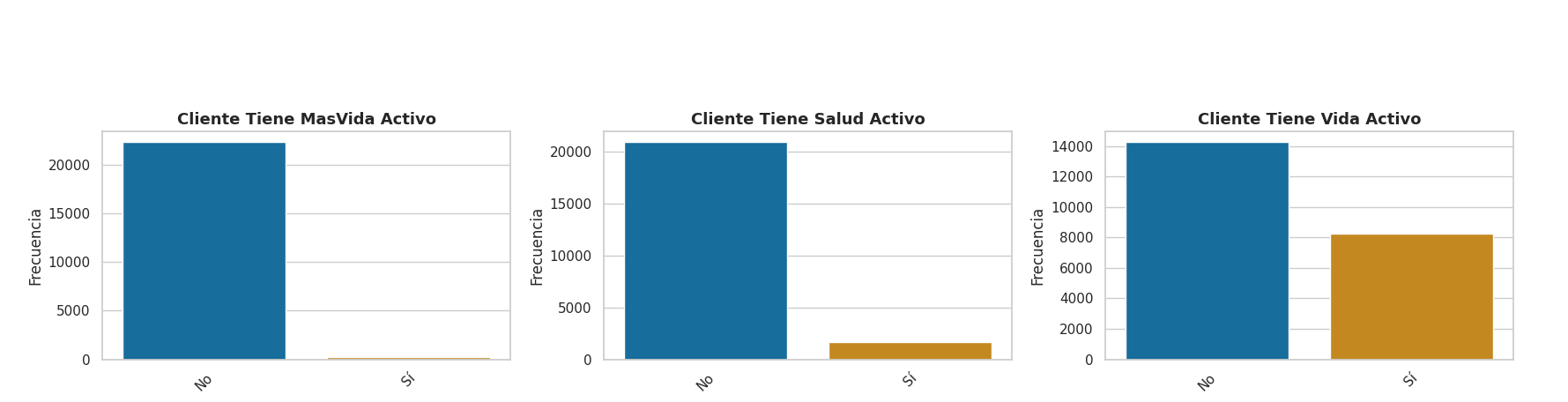
El conjunto de datos final quedó con 37 columnas y 22523 registros

**Validación de Variables Categóricas.**

La mayoría de los clientes pertenecen al segmento "Adulto joven", son de sexo masculino y en su mayoría están ubicados en Antioquia y Bogotá, lo que refleja una alta concentración geográfica y un perfil demográfico definido. La gran mayoría de los vehículos son de uso particular, y se observa un bajo nivel de productos complementarios activos como SOAT o seguros de hogar. Adicionalmente, se identifican problemas de calidad de datos en la variable de estado civil, donde hay alta proporción de registros ‘*sin información*’, lo que podría afectar la precisión de futuros análisis si vamos a tener en cuenta esta variable.



****

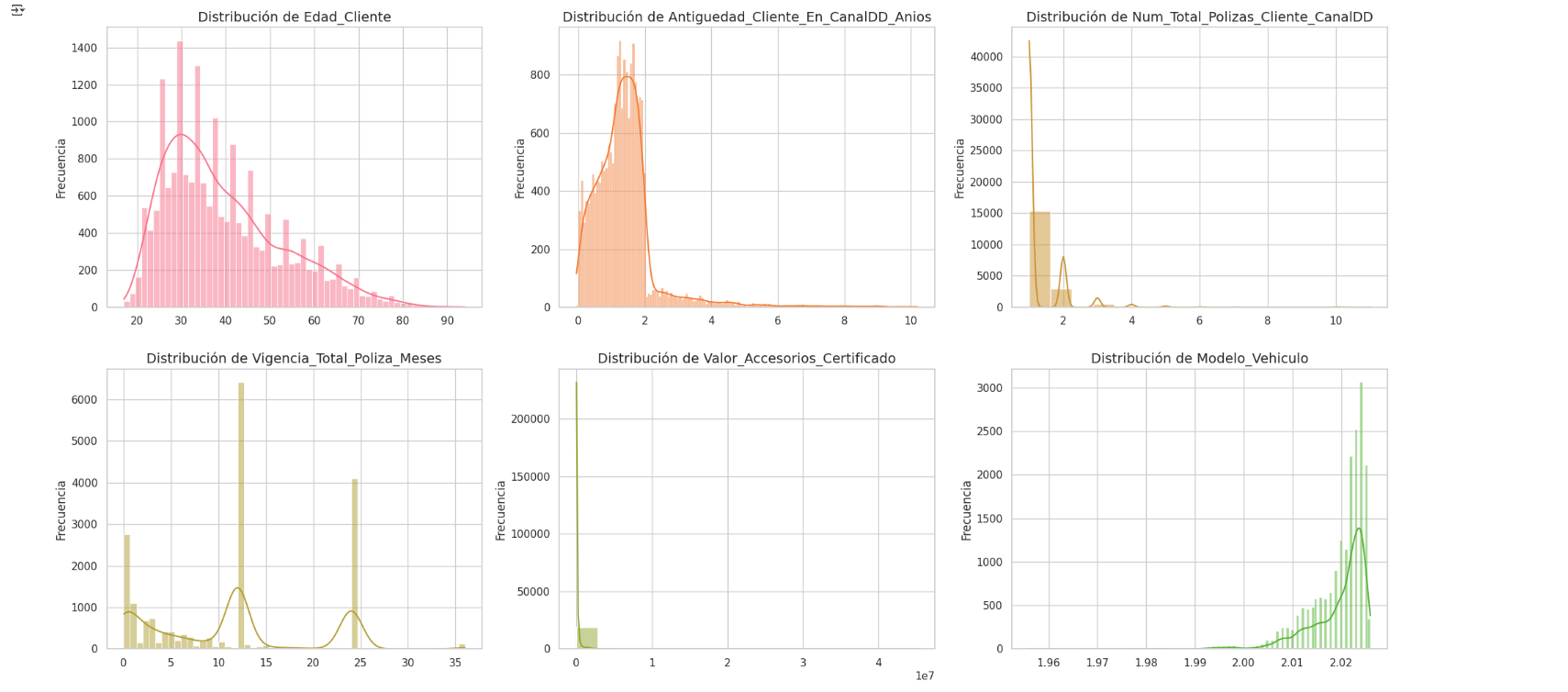
****

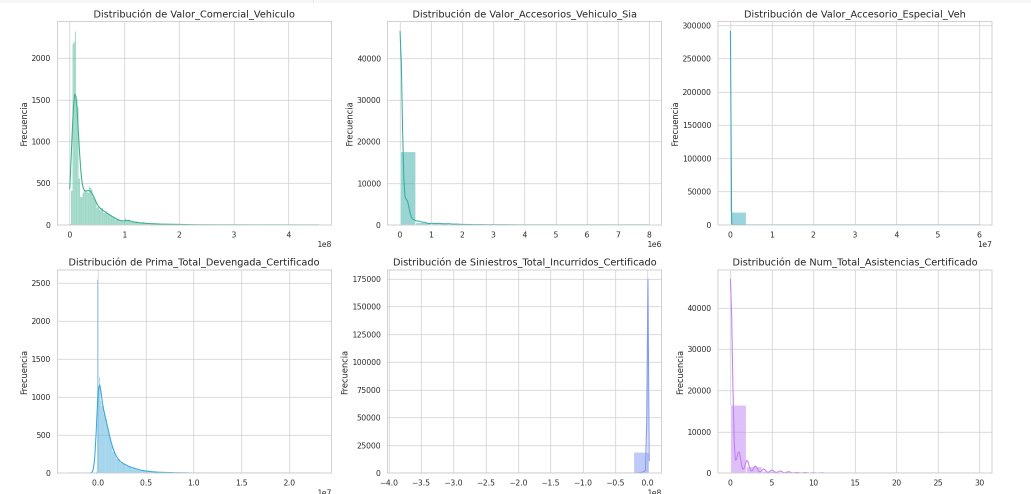
**Validación de variables numéricas**

Las variables numéricas del *dataset* reflejan características clave del perfil del cliente y el estado de sus pólizas. La **Edad del Cliente** y la **Antigüedad en el canal** indican que la mayoría de los clientes son jóvenes, lo que puede influir en su comportamiento y necesidades. El **Número total de pólizas** revela que la mayoría de los clientes tienen pocas pólizas activas, lo cual puede indicar oportunidades para ampliar su portafolio de seguros.

En cuanto a las variables financieras, como el **Valor Comercial del Vehículo**, **Valor de Accesorios**, y la **Prima Total Devengada**, muestran una amplia variabilidad con predominancia de valores bajos y algunos casos de valores muy altos, sugiriendo una heterogeneidad significativa en el tipo de clientes y vehículos asegurados. Esto implica que existen clientes con vehículos de bajo y alto valor.

Las variables relacionadas con siniestros y asistencias indican que la mayoría de los clientes reportan pocas reclamaciones o asistencias, lo cual podría reflejar buen comportamiento o baja frecuencia de eventos adversos, aunque también podría indicar subregistro o clientes con poca experiencia en el uso de sus pólizas.

****

****

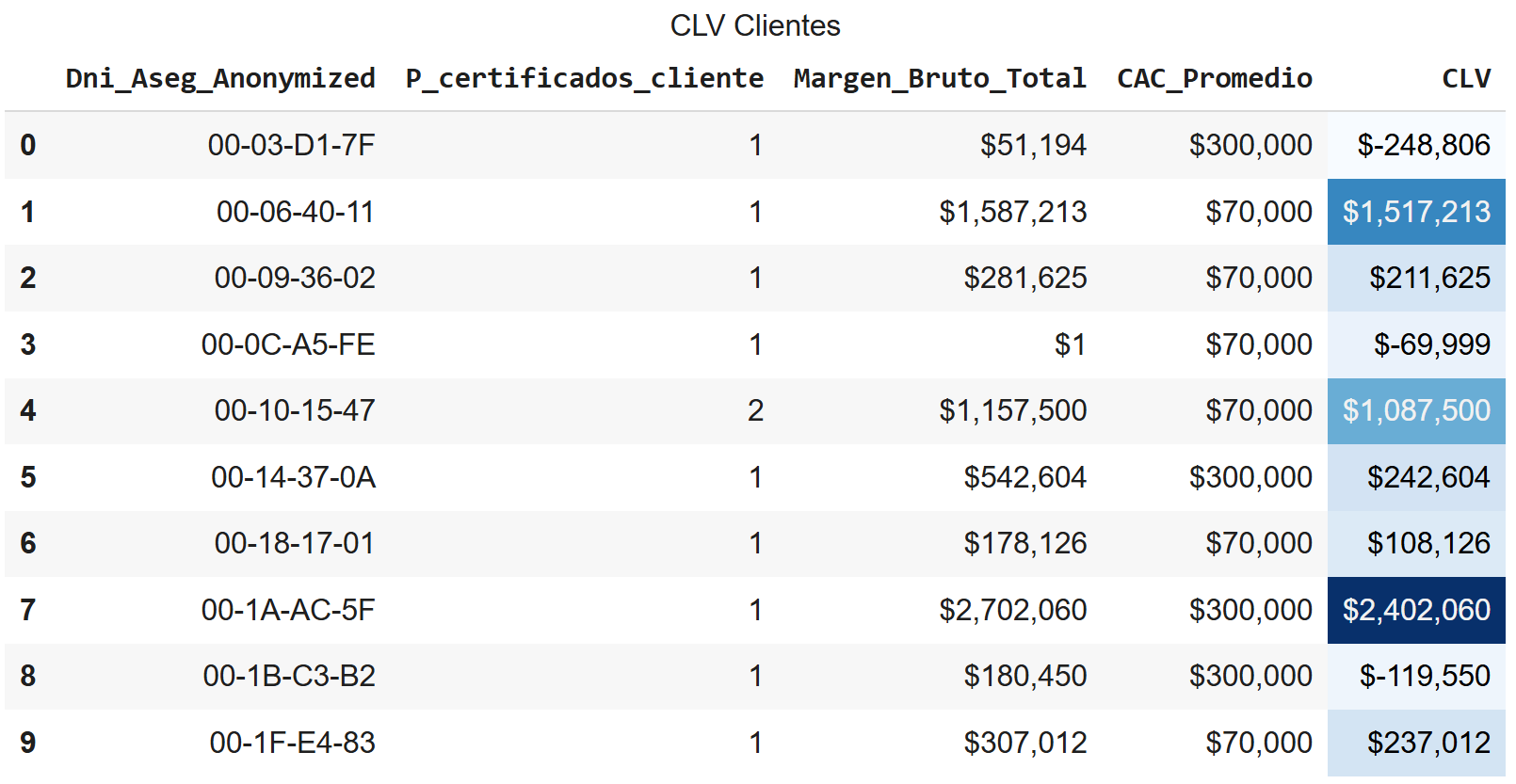
**Inclusión de variables en el Data set**

Dentro de la preparación de los datos debimos incluir dos variables numéricas importantes las cuales nos ayudarán con los diferentes modelos que se van a trabajar, teniendo en cuenta la formula del cálculo del **CLV** propuesto anteriormente.

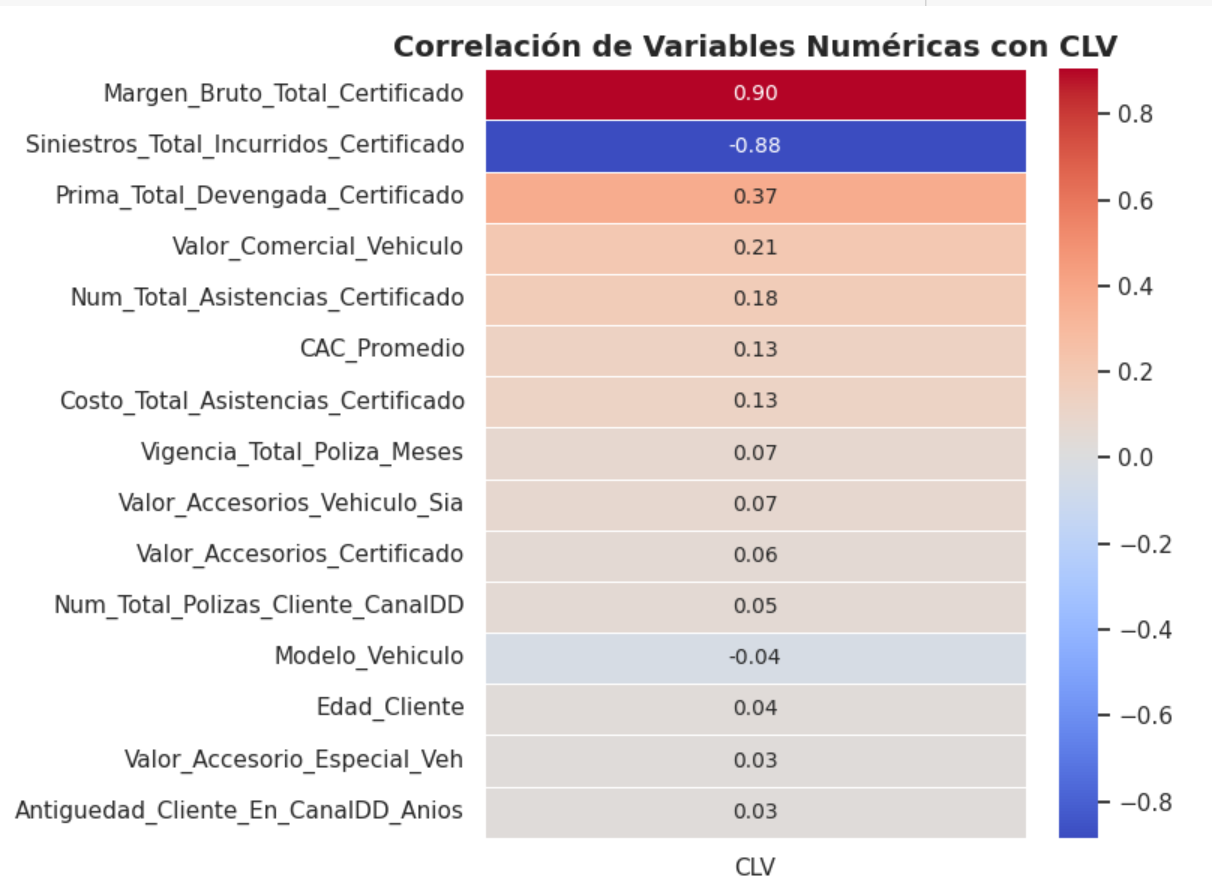
El primero es *‘CAC\_Promedio’* el cual es un dato suministrado por parte de un analista del canal digital, el cual nos informa del *Costo de adquisición promedio* para motos $70.000 y para autos $300.000

También se ingresa la columna llamada ‘*CLV’* el cual contiene la información calculada mediante a formula descrita anteriormente para el hallazgo del *Customer Lifetime Value*.

Es importante resaltar que esta va a ser nuestra Target, importante para la implementación de los modelos.



Una vez validadas todas las variables numéricas y categóricas, y luego de incluir las nuevas 2 columnas (*‘CAC\_Promedio’ y ‘CLV’*) se procedió a realizar de nuevo la matriz de correlación teniendo en cuenta que ya tenemos la variable predictora.



El análisis de correlación muestra que las variables numéricas con mayor relación con el (*Costumer lifetime value*) son el ‘*Margen\_Bruto\_Total\_Certificado’* (correlación positiva de 0.90) y los ‘*Siniestros\_Total\_Incurridos\_Certificado’* (correlación negativa de -0.88). Esto sugiere que un mayor margen bruto está fuertemente asociado con un mayor CLV, mientras que un mayor monto en siniestros incurridos reduce significativamente el CLV. Otras variables como la Prima\_Total\_Devengada\_Certificado (0.37) y el Valor\_Comercial\_Vehículo (0.21) tienen una relación positiva más moderada. Por el contrario, variables como el Modelo\_Vehículo, Edad\_Cliente o Antigüedad\_Cliente\_En\_CanalDD\_Anios muestran una correlación cercana a cero, lo que indica una relación muy débil o inexistente con el CLV.

# Modelado (Propuesto)

[Ingrese la información solicitada aquí sin los corchetes]

# Evaluación (Propuesta)

[Ingrese la información solicitada aquí sin los corchetes]

# Despliegue (Propuesto)

[Ingrese la información solicitada aquí sin los corchetes]

# Referencias

AnderGP512. (2025). *Proyecto-de-Grado* [Repositorio]. GitHub. <https://github.com/AnderGP512/Proyecto-de-Grado>