UNIVERSIDAD
DE SANTIAGO
DE CHILE

UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE CIENCIAS
DEPARTAMENTO DE MATEMÁTICA Y CIENCIA DE LA COMPUTACIÓN

Proyecto de Tesis

Modelación del costo medio de seguros colectivos de salud con herramientas de ciencia de datos y aprendizaje automático

Ingeniería Estadística

Estudiante: Vicente Núñez Roco

Profesor Guia: Francisco Plaza Vega

Comisión: Rodrigo Aroca Gonzales
Omar Chocotea Poca

Fecha: 31 de Marzo, 2025

Santiago de Chile

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

Según la Fundación MAPFRE (2025), un seguro es una actividad económico-financiera que presta el servicio de transformación de los riesgos de diversa naturaleza a que están sometidos los patrimonios, en un gasto periódico presupuestable, que puede ser soportado fácilmente por cada unidad patrimonial. Es decir, un servicio en el cual, a cambio de un pago periódico (prima), se asumen los riesgos asociados al patrimonio asegurado, tal que, en caso de daños al bien, se debe indemnizar el valor de dicho bien.

Help Seguros de Vida S.A. es una empresa aseguradora, parte del holding Banmédica, especializada en productos que brindan cobertura ante siniestros de vida y salud. Sus productos están diseñados para adaptarse a las necesidades específicas de los clientes y pueden ser individuales o colectivos, dependiendo de las características de la entidad contratante. Los seguros colectivos están diseñados para empresas y brindan beneficios a un grupo de trabajadores, sin distinguir características personales.

La tarificación es una actividad encaminada, previos cálculos tectónicos y estadísticos a determinar las tasas o prima aplicable a los diferentes riesgos Fundación MAPFRE (2025). Determinar la prima implica distintos enfoques, dependiendo del tipo de seguro y del contratante. En los seguros de salud, se parte del supuesto de que la cobertura será utilizada durante todo el periodo de vigencia, por lo que es necesario que la prima calculada sea suficiente para cubrir los siniestros esperados. En los seguros colectivos, a diferencia de los individuales, la tarificación no considera características individuales, sino que analiza el comportamiento del grupo asegurado en su conjunto. Además, el proceso implica una negociación con el cliente para que la prima final se ajuste a sus necesidades y presupuesto. En este cálculo, el tarificador considera múltiples variables relevantes, como la edad promedio del grupo, la distribución por género, la siniestralidad histórica y el costo medio por evento. Estos factores influyen directamente en el precio de las coberturas solicitadas.

El costo medio o prima pura, según Fundación MAPFRE (2025), representa la unidad más simple y básica del concepto de prima, por cuanto significa el coste real del riesgo asumido por el asegurador, sin tener en cuenta sus gastos de gestión. Este valor es de gran importancia al momento de tarificar, puesto que, al llegar un nuevo cliente, se desconoce el coste que implica cubrir su riesgo durante la duración de la póliza. Una estimación certera del costo medio permite determinar una prima que minimiza el riesgo de incurrir en pérdidas sin dejar de tener un precio competitivo.

Este valor surge como resultado de multiplicar dos índices que describen la utilización del seguro: la frecuencia, que según Fundación MAPFRE (2025) corresponde a un coeficiente que refleja el promedio del número de siniestros que una póliza de seguros tiene durante un año completo, y que en seguros colectivos se explica como la cantidad de siniestros por sujeto asegurado durante un periodo de tiempo. Según Frees (2018), la severidad denota la cantidad o el tamaño de cada pago por un evento asegurado, es decir, la magnitud del uso del seguro, que en seguros colectivos representa el costo promedio de cada siniestro.

Help Seguros de Vida S.A. actualmente carece de una metodología robusta para la modelación del costo medio y, además, la cartera de seguros no es lo suficientemente grande para realizar cálculos empíricos confiables. Lo anterior produce problemas como la subestimación de la prima y, como consecuencia, que dicha prima no logre cubrir los gastos esperados; o una sobreestimación de la misma, que aumenta los precios por sobre el mercado, haciendo necesario recurrir a otras herramientas para reducir la tarifa final.

Debido a esto, surge la necesidad de desarrollar una metodología acorde a esta información para modelar el costo medio de los distintos servicios que reciben reembolso por parte de la compañía, mediante el uso de modelos estadísticos y machine learning, utilizando el lenguaje de programación R, con el objetivo final de contribuir a la mejora de la tarificación de los productos.

A partir de lo anterior, el proyecto consiste en desarrollar una metodología acorde a la información disponible para la modelación estadística del costo medio, enfocada en la tarificación de pólizas colectivas de salud.

1.1. Objetivos del Proyecto

1.1.1. Objetivo General

- Desarrollar una metodología para la modelación estadística del costo medio para la compañía aseguradora Help Seguros de vida S.A., considerando las características propias de la información disponible.

1.1.2. Objetivos Específicos

- Realizar una fase de recopilación y preparación de la información disponible.
- Realizar un análisis exploratorio de las variables obtenidas, para identificar patrones y asociaciones en los datos.
- Implementar distintos modelos estadísticos y de machine learning (Utilizando R), que permitan explicar la frecuencia y la severidad.
- Evaluación de los modelos propuestos.

1.2. Estado del Arte

El problema de modelar los costos medios es uno que se observa frecuentemente en las compañías de seguros, tanto de salud, automotriz como de vida. A continuación, se presentan algunas aproximaciones recientes para abordar este problema, bajo los distintos contextos:

- Zhang et al. (2012): Bajo el contexto de seguros de accidentes laborales, plantean un acercamiento basado en un modelo jerárquico bayesiano, obteniendo resultados positivos para la estimación y predicción de los costos. No es posible aplicarlo debido a que la estructura de la información que plantean es muy distinta a la disponible.

- Tuininga (2022): Bajo el contexto de seguros automotrices, compara la eficiencia de modelos lineales generalizados (GLM) y modelos de machine learning para la estimación de la frecuencia y severidad, concluyendo que los modelos GLM tienen mejor ajuste. Plantea problemas a futuro como el estudio de variables a utilizar y el tratamiento de una base de datos desbalanceada.

- Guelman (2012): Plantea el modelo Gradient Boosting Trees como alternativa a los modelos lineales generalizados para modelar la frecuencia y la severidad. Como resultado, se obtiene que los modelos lineales generalizados siguen teniendo un mejor desempeño para este tipo de problemas. Se diferencia del problema planteado debido a su enfoque en seguros individuales y a que el conjunto de variables utilizadas no es mostrado.

- Poufinas et al. (2023): Bajo el contexto de seguros automotrices, plantea el uso de modelos de machine learning para los reclamos, donde a partir de una configuración de la base de datos propone una metodología para realizar predicción. A diferencia del proyecto actual, se debe estudiar si es posible replicar el uso de dicha configuración ajustada al contexto de seguros colectivos.

- Graziadei et al. (2023): Plantea el modelamiento de la frecuencia y la severidad para seguros de automóviles, mediante Random Forest y modelos de regresión generalizados para cada variable. Modela la frecuencia con un modelo Poisson y la severidad con un modelo de regresión Gamma. Finalmente, compara el desempeño de ambos modelos, siendo Random Forest el que obtuvo mejor resultado. Sin embargo, sus resultados no son totalmente aplicables, debido a que el conjunto de variables usadas por el autor es muy reducido y toma la frecuencia y la severidad durante todo el contrato.

- Piontkowski (2020): Plantea el ajuste y la proyección de los costos medios mediante modelos estocásticos y series de tiempo tipo ARIMA. A diferencia del problema planteado, se dispone de una cartera lo suficientemente longeva como para poder realizar modelos muy certeros.

- García et al. (2023): Compara métodos de tarificación centrados en la teoría de la credibilidad con el modelo financiero Black-Scholes, concluyendo que los modelos basados en la teoría de la credibilidad tienen mejor desempeño. A diferencia del proyecto planteado, propone ajustar las variables a distintas distribuciones para evaluar el riesgo de asegurar cierto grupo.

- Cordeiro (2023): Plantea un estudio como solución a un problema de renovación de tarifas para seguros colectivos, en el cual explican que para ello es necesario realizar proyecciones de los últimos tres meses de costos. Para lo cual propone distintos modelos de machine learning como alternativa a modelos de series temporales. En los meses proyectados obtuvo mejores resultados que el modelo utilizado anteriormente. El proyecto no hace mención de la configuración de la base de datos ni de las variables utilizadas, por lo tanto, se desconoce si se puede replicar la misma solución.

Capítulo 2

Marco Teórico

2.1. Conceptos Clave

2.1.1. Prima Pura o Costo Medio

Como se mencionó en la seccion 1, la prima pura representa la unidad más simple y básica del concepto de prima, ya que significa el costo real del riesgo asumido por el asegurador, sin tener en cuenta sus gastos de gestión. Este valor es importante, puesto que depende del riesgo que se está asegurando y, a diferencia de otros productos, al momento de vender se desconocen los costos de proteger el patrimonio. Debido a esto, la fijación de precios difiere de la de otros productos financieros. Según Frees (2018), el seguro implica una promesa de la aseguradora de pagar un reclamo cuando el asegurado lo presente. La prima pura es usada en el enfoque técnico de la fijación de precios, donde la tarifa final o prima final se calcula de la siguiente forma:

$$Prima = \frac{Prima\,pura + GastosFijos}{1 - Tasa\,de\,Gastos\,Variables - Ganacias} \tag{1}$$

Donde la prima pura se calcula de la siguiente manera:

$$Prima\,Pura = \frac{Recuento\,de\,Reclamos}{Expuestos} * \frac{Gasotos\,de\,Reclamos}{Recuento\,de\,Reclamos} \tag{2}$$

Lo anterior se denomina la ecuación de prima simplificada, puesto que no considera la inversión inicial y combina la consideración de los reclamos y la magnitud de estos, además de incluir un término llamado expuestos. Esta es una variable que refleja el riesgo que está por aceptar la compañía. Este puede ser representado por el periodo durante el cual el contrato esté vigente, y en el contexto de seguros colectivos, corresponde a la población expuesta en un periodo de tiempo. Estos dos componentes fueron nombrados en 1 como la frecuencia y la severidad de los siniestros.

- **Frecuencia**: Según Frees (2018), es la frecuencia con que ocurre un evento asegurado dentro de un contrato de póliza, es decir, el número de reclamaciones. El estudio de esta variable tiene importancia a nivel contractual, donde, según Frees (2018), los datos de conteo que representa esta variable indican un factor importante al momento de fijar el valor del deducible o los límites de la póliza.
  A nivel conductual, se estudian condiciones que minimizan la utilización del seguro. En el caso de seguros individuales, a través del cuidado preventivo y mediante filtros mediante las características de los asegurados; en cambio, en seguros colectivos, es mediante características que cumpla el grupo.
  Según Frees (2018), para modelar la frecuencia, se utilizan las siguientes distribuciones: Binomial, Poisson y Binomial Negativa.

- **Severidad**:Según Frees (2018), denota la cantidad o tamaño de cada pago por un evento asegurado.

Analizar los cambios en la severidad proporciona información sobre las tendencias de pérdidas y resalta el impacto de cualquier cambio en los procedimientos de manejo de reclamaciones. Por lo tanto, su estudio debe realizarse en conjunto con el de la frecuencia. Según Frees (2018), se estudia bajo las siguientes distribuciones: Gamma, Pareto, Weibull, Tweedie.

2.2. Herramientas Estadísticas

2.2.1. Test de Kruskal-Wallis

Considerando una colección de k muestras aleatorias independientes. El análisis de varianza unifactorial de Kruskal-Wallis es una prueba que permite decidir si k muestras independientes provienen de diferentes poblaciones, es decir, si los valores de las muestras difieren y si estas diferencias representan variaciones que pueden esperarse en muestras que se obtienen al azar de una misma población.

Donde la hipótesis nula indica que las k muestras provienen de una misma población con una misma mediana, es decir:

Hipótesis:

$$H_0 : \theta_1 = ... = \theta_i = ... = \theta_k \qquad v/s \qquad H_1 : \exists\ j,\ j \neq i,\ \text{tal que},\ \theta_i \neq \theta_j$$

Donde θ_i corresponde a la mediana de la muestra i y la hipótesis alternativa indica que al menos una de las muestras o grupos tiene una media diferente.

Procedimiento

1. Los datos se ordenan en la siguiente tabla: Donde:

Grupos				
1	...	j	...	k
X_{11}		X_{1j}		X_{1k}
		...		
X_{i1}	...	X_{ij}	...	X_{ik}
		...		
$x_{n_1 1}$	...	$X_{n_j j}$	...	$X_{n_k k}$

   - n_j es el tamaño del grupo j.
   - X_{ij} es la i-ésima observación del j-ésimo grupo.
2. Se ordenan los datos de las k muestras en una sola muestra.
3. Se les asignan rangos (r_{ij}) de forma ascendente a la muestra combinada.
4. Se calcula la suma de rangos de cada grupo como $R_j = \sum^{n_j} i = 1rij$.

Estadístico: Si las muestras provienen de una misma población, se espera que los rangos promedios de cada grupo deberían ser aproximadamente los mismos. Para verificar lo anterior, se utiliza el siguiente estadístico:

$$KW = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (\bar{R_j} - \bar{R})^2 = \left[\frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \bar{R_j}^2 \right] - 3(N+1)$$

Donde:

- $N = \sum_{j=1}^{k} n_j$: Total de observaciones.
- $\bar{R}_j$: Promedio de los rangos de la j-ésima muestra o grupo.
- $\bar{R} = \frac{N(N+1)}{2}$: Promedio de los rangos de la muestra combinada.

Región de Rechazo:

$$RC = \text{m.a.}(N) \ / \ KW^{\text{obs}} \geq \chi^2(k-1)(1-\alpha)$$

2.2.2. Correlación de Pearson

Corresponde a una medida estadística que, según Montgomery y Runger (2014), se define como una medida de la intensidad y relación lineal entre 2 variables aleatorias continuas y se calcula de la siguiente forma:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

donde:

- x_i y y_i son los valores de las variables en estudio.
- $\bar{x}$ y $\bar{y}$ son las medias de las variables x e y, respectivamente.
- n es el número de observaciones.

La interpretación de r depende de su magnitud y signo:

- Valores cercanos a 1 o -1 indican relaciones lineales fuertes.
- Valores cercanos a 0 indican relaciones lineales débiles o inexistentes.

2.2.3. Componentes Principales

El análisis de Componentes Principales es utilizado para el análisis de datos no agrupados y consiste en generar nuevas variables llamadas componentes, a partir de combinaciones lineales de un conjunto de variables que presentan correlación entre sí, tal que cada componente capte la mayor cantidad de variabilidad del conjunto original, siendo estos ortogonales entre sí. Esta técnica de reducción de dimensionalidad es útil para poder utilizar conjuntos de variables que presenten dependencia, como en modelos de regresión u otras técnicas que, entre sus supuestos, requieren que el conjunto de variables predictoras sea independiente entre sí. Para el caso de este estudio, se desconoce si existe dependencia entre las variables a estudiar, pero debido a la limitada cantidad de variables, es necesario el uso de estos análisis para evitar perder la mayor cantidad de información posible. A continuación, se da una explicación de la obtención de los componentes:

- **Primera Componente Principal**: Corresponde a la combinación lineal de variables con varianza máxima. Entonces, dado el conjunto de variables $X = (\vec{x_1}, ..., \vec{X_n})$, la primera componente está dada por:

  $$Z_1 = Xa \quad \text{Max } \mathbf{V}(Z_1) = \frac{1}{n} Z_1^t Z_1 = \frac{1}{n} a^t X^t X a = a^t S a$$

  Donde a es el vector que contiene la transformación lineal y S es la matriz de varianza-covarianza. El objetivo es maximizar la varianza, para lo cual se debe obtener el a óptimo bajo la siguiente restricción $a^t a$, que se introduce mediante multiplicadores de Lagrange de la siguiente manera:

  $$M = a^t S a - \lambda(a^t a - 1)$$

Minimizándola respecto a a e igualándola a 0, se obtiene que:

$$|S - I\lambda|a = 0$$

De lo cual se concluye que λ tiene que ser igual a la varianza de Z, y como lo que se busca es maximizar, entonces λ corresponde al mayor valor propio de S y a es el vector asociado.

- **Segunda Componente**: Para el cálculo de la segunda componente, se busca maximizar la varianza de la suma de $Z_1 = Xa_1$ y $Z_2 = Xa_2$, tal que a_1 y a_2 sean ortogonales, para que los componentes no tengan correlación. Para lo cual se impondrán más restricciones al problema de maximización. Entonces, el problema queda de la siguiente manera:

$$\begin{cases} \mathbf{a}_1^t \mathbf{a}_1 = 1, \\ \mathbf{a}_2^t \mathbf{a}_2 = 1, \\ \mathbf{a}_1^t \mathbf{a}_2 = 0, \end{cases}$$

  De lo cual, derivando respecto a a_1 y a_2 y resolviendo el sistema de ecuaciones, se obtiene que λ_i son los dos valores propios de mayor magnitud de S y que a_i son los vectores propios asociados.

- **Generalización de los Componentes**: En general, la matriz X tiene rango p y, en consecuencia, p componentes principales. Las cuales se obtienen calculando los valores propios de la matriz S de varianzas-covarianzas, mediante la siguiente ecuación:

$$|S - \lambda_i I|a_i = 0$$

  Y debido a que $a_i \neq 0$, entonces se deberá resolver la ecuación, así obteniendo valores propios de los cuales se selecciona el i-ésimo máximo valor propio, que está asociado al i-ésimo vector propio, que será el valor del vector a_i.

Decidir cantidad de Componentes principales En la aplicación, se debe decidir la cantidad de componentes a utilizar, con el objetivo de reducir la dimensionalidad del conjunto original de tal manera que sea mínima la pérdida de información.

- Rencher (2002) recomienda retener la cantidad de componentes, tal que la cantidad de variabilidad capturada sea del 80 %.
- Retener los componentes propios, tal que su valor propio sea mayor a la media de los valores propios de la matriz S.

2.3. Propuestas de Modelos

En la siguiente sección se presentan los modelos propuestos para el tratamiento de las variables a modelar, elegidos debido a la influencia del estado del arte y para cubrir las características de entrada de cada problema que se requiera solucionar.

2.3.1. Modelos Lineales Generalizados GLM

Los modelos lineales generalizados (GLM) son herramientas estadísticas que permiten el estudio del comportamiento de una variable respuesta Y respecto a un conjunto de p variables predictivas $(X_1, \ldots, X_p)$, donde la variable respuesta pertenece a la familia exponencial, que según lo observado en la sección 2.1.1

corresponde a los casos de estudio donde se suele utilizar la distribución normal y poisson. Estos modelos se componen de tres componentes, los cuales son nombrados a continuación:

- Componente aleatorio: Corresponde a la variable respuesta $Y = (y_1, \ldots, y_n)$, la cual tiene densidad de probabilidad en la familia exponencial.
- Componente sistemático o productor lineal: Corresponde a la combinación lineal de variables explicativas que permite explicar la variable respuesta, esta está dada por:

$$\eta_i = X^T\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, \ldots, n$$

- Función de enlace $g(\cdot)$: Corresponde a la función que conecta el componente aleatorio con el componente sistemático, la cual esta especificada por la distribución del componente aleatorio. Si definimos como $\mu_i = \mathbb{E}(y_i)$, entonces la función de enlace relaciona los componentes del modelo de la siguiente forma:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, \ldots, n$$

La cual puede tomar las siguiente formas dependiendo de la función de enlace a utilizar:

Tabla 1: Tabla Modelos Lineales Generalizados

Distribución	**Función de enlace**	**Modelo final**	**Objetivo**
Normal	Identidad: $g(\mu) = \mu$	$\mu = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$	Modelar variable continua
Binomial	Logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$	Regresión logística
Poisson	Log: $g(\mu) = \log(\mu)$	$\log(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$	Modelar conteo

2.3.2. Árboles

Los árboles de regresión o clasificación son algoritmos que identifican formas de dividir un conjunto de datos en segmentos homogéneos llamados hojas, partiendo de un nodo raíz, el cual representa la división en la variable predictora. El algoritmo decide cuál es el mejor punto de división minimizando una función de pérdida. Este proceso es iterativo, donde en cada paso, cada nuevo subconjunto se divide en nuevos nodos hasta alcanzar un criterio de parada. El resultado entregado por el árbol dependerá del objetivo que se tenga de este:

1. **Regresión**: Si el problema a tratar es de regresión, dadas las particiones del espacio $R_1, R_2, \ldots, R_j$, y sea c_j el valor característico en cada región, entonces la función que identifica este valor está dada por:

$$f(x) = \sum_{j=1}^{J} c_j I(x \in R_j)$$

Dado lo anterior, el valor característico debe ser estimado para cada región, que bajo la función de pérdida cuadrática $\sum(y_i - f(x_i))^2$, el mejor $\hat{c_j}$ está dado por:

$$\hat{c_j} = \frac{\sum(y_i | x_i \in R_j)}{n},$$

2. **Clasificación**: En este caso, para $\hat{p}_{jk}$ sea la proporción de la clase k en el nodo j para N_j observaciones, tenemos:

$$\hat{p}_{jk} = \frac{1}{N_j} \sum_{x_i \in R_j} I(y_i = k)$$

Entonces se define el resultado del nodo j como la clase tal que $k(m) = \arg\max_k \hat{p}_{jk}$, es decir, la clase que tenga la mayor proporción dentro de la región. Para determinar medidas del error de clasificación o impureza de los nodos, se utilizan las siguientes fórmulas:

- **Índice de Gini**: $\sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$
- **Devianza**: $-\sum_{k=1}^{K} \hat{p}_{jk} \log \hat{p}_{jk}$

La desventaja del modelo es el sobreajuste, el cual ocurre cuando el árbol se hace más complejo que su tamaño óptimo, provocando que sea muy susceptible a cambios muy pequeños de nuevas observaciones, por lo tanto, reduciendo la potencia de predecir nuevas observaciones.

2.3.3. Métodos de Boosting

Corresponden a métodos que funcionan mediante la suma de múltiples árboles, de la siguiente forma:

$$f_M(x) = \sum_{m=1}^{M} T(x; \theta_m)$$

Donde $T(x; \theta_m)$ corresponde a un árbol de regresión. En cada iteración del procedimiento se debe obtener las regiones R_j de cada árbol, para lo cual se debe resolver:

$$\hat{\theta} = \arg\min_{\theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m))$$

Para el conjunto de regiones y constantes $\theta_m = \{R_{jm}, c_{jm}\}_1^{J_m}$ y dado el modelo anterior $f_{m-1}(x)$. Entonces, una vez definido R_{jm}, se procede a optimizar las constantes c_{jm}, que para cada región se obtienen mediante:

$$\hat{c_{jm}} = \arg\min_{c_{jm}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + c_{jm})$$

2.3.4. Random Forest

Introducido por (Breiman, 2001), corresponde a un modelo de regresión o clasificación basado en árboles, que surge como extensión del bagging y es competidor de los modelos de boosting. Donde para un vector aleatorio $\vec{X} = (X_1, ..., X_p)^T$, que representa la matriz de variables predictoras, y una variable aleatoria Y que representa la variable respuesta, se asume la existencia de una distribución conjunta $P_{XY}(X, Y)$ y se establece el objetivo de encontrar una función $f(X)$ que prediga Y, determinada por una función de pérdida $L(Y, f(X))$, que mide qué tan alejada está $f(X)$ de Y. Para ello, se minimiza la esperanza de la función de pérdida:

$$E_{XY} L(Y, f(X))$$

La construcción de $f(X)$ se basa en la técnica de aprendizaje en conjunto, que utiliza una colección de modelos base $(h_1(X), ..., h_j(X))$. Estos modelos combinados generan la predicción conjunta: en problemas de regresión, los resultados se promedian, mientras que en clasificación se elige la clase con mayor frecuencia.

En el modelo Random Forest, los modelos base son los árboles descritos en 2.3.2. Cada árbol se denota como $h_j(X, \Theta_j)$, donde $\Theta_j = (\theta_1, ..., \theta_j)$ es un vector aleatorio que introduce aleatoriedad en el modelo de dos maneras:

1. **Mediante el bagging**: Técnica en la que cada modelo base se entrena con una muestra bootstrap de la base original. En este caso, θ_j está involucrado en la inicialización de la muestra.
2. **En la selección de variables para la división de nodos**: Al dividir un nodo, la mejor división se determina a partir de una muestra aleatoria de tamaño m de las variables predictoras, donde cada muestra se selecciona de manera independiente en cada nodo.

El algoritmo resumido puede ser visto en A.2.

Predicciones con datos Out-of-Bag: Al momento de tomar la muestra bootstrap, no todas las observaciones entran en el set de entrenamiento. Estas observaciones son las Out-of-Bag, y son necesarias para realizar una correcta estimación del error del modelo con los datos de testeo y determinar la importancia de las variables. Para determinar el error, para cada dato en el set de entrenamiento, se utilizarán árboles que tengan esta observación como un Out-of-Bag. Este error, denominado error de generalización, es estimado mediante el error cuadrático medio (MSE) de la siguiente forma para un problema de regresión:

$$MSE_{\text{oob}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}_{\text{oob}}(x_i))^2 \tag{3}$$

Y para un problema de clasificación:

$$E_{\text{oob}} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{f}_{\text{oob}}(x_i)) \tag{4}$$

Importancia de las Características: La importancia de las características en Random Forestçorresponde al peso que tienen las variables de estudio en el conjunto de árboles. Esta cobra importancia al momento de seleccionar las variables, para evitar variables redundantes y usar solo los predictores que tienen mayor impacto en el modelo. Durante el presente trabajo se utilizará la técnica de importancia por permutación, donde el procedimiento consiste en:

1. Obtener las predicciones out-of-bag del modelo.
2. Repetir lo siguiente para cada variable predictora $k = 1, \ldots, j$:
3. Los valores de la variable k son permutados aleatoriamente, manteniendo el resto de variables constantes.
4. Se vuelve a calcular las predicciones out-of-bag.
5. Para cada predicción real y permutada se calcula la diferencia.

Mediante este procedimiento se da una medición del impacto que tiene el valor de la variable k dentro del modelo.

2.3.5. Redes neuronales

Las redes neuronales profundas son modelos de aprendizaje automático, cuyo objetivo es aproximar una función $f^*(x)$, mediante la combinación de múltiples funciones, donde la función aprende qué vector de parámetros β resulta en dar la mejor aproximación. Este aprendizaje se realiza desde la evaluación de la

matriz de variables predictoras $\vec{X} = (X_1, \ldots, X_p)^T$, pasando por cálculos intermedios con el fin de definir f y obtener una salida y.

Una característica de estos modelos es la capacidad de poder capturar relaciones no lineales, lo que diferencia de los modelos tradicionales (GLM) es la estructura que utiliza el modelo, puesto que se compone de tres tipos de capas: la capa de entrada, las capas ocultas y de salida.
La capa de entrada es la capa inicial del modelo donde se reciben las variables predictoras $\vec{X} = (X_1, \ldots, X_p)^T$, llamadas características. Las capas ocultas corresponden a la etapa del modelo donde se procesa la información de entrada y se aplican transformaciones no lineales con el objetivo de capturar relaciones complejas que las regresiones lineales no podrían. Cada capa oculta se estructura de K unidades llamadas neuronas, las cuales tienen la siguiente forma:

$$A_k = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j\right)$$

Donde, $g(*)$ es una función de activación no lineal, la cual es definida según el tipo de variable que se necesite estudiar. Esto se hace secuencialmente para cada unidad, las cuales alimentan la capa de salida, que funciona de manera similar a una capa oculta pero utilizando una función de activación correspondiente al problema, obteniendo el siguiente modelo:

$$f(x) = g\left(\beta_0 + \sum_{k=1}^{K} \beta_k A_k\right)$$

Las redes neuronales suelen tener más de una capa oculta, donde cada capa $l = 1, \ldots, L$ toma las activaciones A_k de la capa anterior como nuevas entradas, formando una red de transformaciones, capaz de construir relaciones complejas de datos de la siguiente forma:

$$A_m^{(l)} = g(w_{m0}^{(l)} + \sum_{j=1}^{J} w_{mj}^{(l)} A_j^{(l-1)})$$

Función de Activación: Corresponde a una función que realiza una transformación sobre los valores de la entrada de x, que según (Goodfellow et al., 2016), corresponde a una función no lineal y cuya función es permitir a la red capturar relaciones complejas. A continuación, se presentan las funciones más comunes, utilizadas en el estudio:

- **Rectified linear unit(relu):** $g(x) = max0, x$
- **Gaussian error linear unit(gelu):** $g(x) \approx 0.5x(1 = tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3))$
- **Sigmoide:** $g(x) = \frac{1}{(1+e^{-x})}$
- **Linear:** $g(x) = x$
- **Exponencial:** $g(x) = e^x$

Función de Costo: Corresponden a funciones utilizadas para optimizar los parámetros del modelo, estas se estudian como la diferencia entre los valores reales y los predichos con la salida del modelo. El objetivo del modelo es poder minimizar estas funciones. La elección de una función de pérdida depende del modelo y la arquitectura que se usará. A continuación, se presentan las más comunes:

- **Mean Squared Error:** Corresponde al promedio de las diferencias cuadráticas entre los valores predichos y los reales

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Mean Absolute Error:** Se define como el promedio de las diferencias absolutas entre los valores reales y predichos.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Cross Entropy:** También conocido como log-loss, es utilizado en problemas de clasificación binaria. Este mide la diferencia entre la probabilidad predicha y la etiqueta real:

$$L(y, p) = -\left(y \log(p) + (1 - y) \log(1 - p)\right)$$

- **Huber Loss:** Combina las propiedades del MAE y el MSE, usado para regresiones robustas:

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta, \\ \delta\left(|y - \hat{y}| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases}$$

- **Poisson Loss:** Usado cuando la variable representa datos de conteo y se asume distribución Poisson:

$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i \log(\hat{y}_i)\right),$$

Capítulo 3

Metodología

3.1. Desarrollo Base de Datos

El primer objetivo de la tesis es realizar una base de datos con la información disponible que representa la realidad de la compañía y nos permita desarrollar modelos estadísticos y de machine learning. La base de datos se realizará en las siguientes etapas:

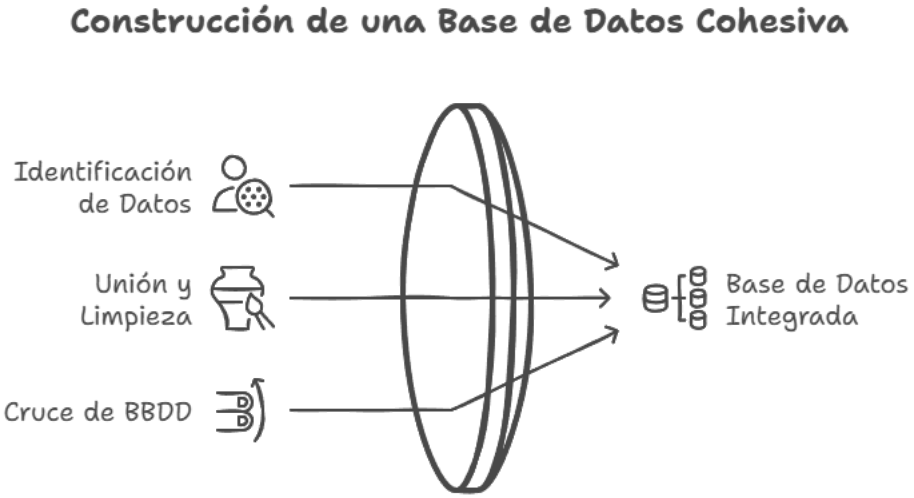

Figura 1: Metodología BBDD

3.1.1. Identificación y Recolección de Datos

Durante esta sección se describirán las fuentes y las variables que se utilizarán al momento de realizar modelos.

Información de los Siniestros: Corresponde a toda la información que está relacionada con los siniestros aceptados por la compañía y que obtuvieron su respectivo reembolso. Estas se extraen de un único archivo que contiene información asociada a los siniestros, llamado sabana de gastos, el cual se actualiza mensualmente y contiene información histórica de los siniestros. Esta base contiene variables de interés que identifican al siniestro, tales como:

Tabla 1: Clasificación de las variables de los siniestros

Identificación	**Costo**	**Características**
Póliza a la que pertenece	Valor de la prestación	Servicio prestado
Rut del asegurado	Abono de la Primera capa	Primera capa del asegurado
Periodo de pago del siniestro	Valor del reembolso	Prestador de salud
	Valor del deducible aplicado	Tipo de póliza
		Razón social del contratante.

Información del Grupo Asegurado : Corresponde a toda la información que permita caracterizar al grupo, mediante información básica de cada asegurado, la cual son obtenidas de los registros de los pagos realizados. Estos archivos se actualizan mensualmente para cada póliza. Este archivo recopila variables como:

- Rut del asegurado.
- Relación con el titular.
- Género del asegurado.

Información de la Cartera : Corresponde a variables que permitan conocer el estado de la cartera de la empresa. Esta puede ser obtenida en el archivo de primas, que recoge información sobre los pagos y los movimientos de las pólizas a lo largo del tiempo. Este archivo se actualiza mensualmente. Recoge variables como:

- Rut del contratante, necesario para realizar futuros cruces de información.
- Evolución de titulares de la póliza.
- Evolución de cargas de la póliza.

Información Geográfica : Corresponde a variables que determinan la ubicación geográfica de los centros de salud, de los cuales se han registrado siniestros, con el objetivo de encontrar las zonas más críticas y de mayor impacto en las variables que se estudiarán. Este conjunto de variables se puede encontrar a través de información pública proveída por el servicio de impuestos internos (SII), la cual puede consultarse en (SII) (2024). De esta base se extraen variables como:

- Rut del Prestador de Salud
- Región del Prestador de Salud
- Comuna del Prestador de Salud

3.1.2. Unión y Limpieza de Datos

Esta parte del proceso está enfocada en realizar la unión de todas las bases que tienen un mismo origen, pero por su naturaleza están dispersas en múltiples archivos, como las bases de cobranza, y realizar una limpieza efectiva de las bases obtenidas previamente. El enfoque está en desarrollar inicialmente los filtros para que la base reflejen lo que se quiere estudiar, que en este caso es tener siniestros que fueron pagados por la compañía y que estén asociados a consultas médicas ambulatorias. Luego de realizar los filtros, el

objetivo es analizar todas las variables cuantitativas para verificar que la información esté bien procesada para futuros moldeamientos. En cuanto a las cualitativas, se recategorizan con el objetivo de eliminar categorías redundantes y revisar categorías cuyo peso en el estudio es mínimo debido a su cantidad de observaciones.

3.1.3. Cruce de Bases de Datos

Durante esta sección se describirán los cruces necesarios para determinar la base de datos final, indicando las variables que se usarán para construir relaciones uno a uno. Estos cruces se realizan en base a la sábana de siniestros, la cual se utilizará como base para los cruces:

- **Cruce con la Base del SII**: Se realizará, usando como identificador, el Rut del prestador de salud o su razón social.
- **Cruce con base de Expuestos**: Se realizará, usando como identificador, el periodo de tiempo.
- **Cruce con Base Cobranza**: Se realizará, usando como identificador, el Rut del asegurado como también la póliza.

Ya con la base de datos limpia y con todas las variables de estudio, se procederá a generar dos bases de datos. La primera, con el objetivo de modelar el reembolso aplicado, se utilizará la misma base, mientras que la segunda estará enfocada en modelar el número de siniestros. Debido a que es una variable que no existe inicialmente, se genera reduciendo la base según la cantidad de agrupaciones formadas por las variables cualitativas, finalmente, obteniendo una variable de conteo que permita modelar la frecuencia.

3.2. Análisis Descriptivo

En esta sección se describirán el procedimiento seguido para desarrollar el análisis descriptivo. Se detallará la estructura utilizada para generar los gráficos, incluyendo las transformaciones necesarias para visualizar la información de manera efectiva. Además, se presentarán las estrategias utilizadas para identificar patrones y relaciones entre variables que puedan ser relevantes para la aplicación de técnicas y/o modelos estadísticos en etapas posteriores. Para lo cual, se seguirá la siguiente estructura:

1. Inicialmente se estudiará la proporción de observaciones de cada variable, con el objetivo de identificar las categorías menos representadas.
2. Se estudiará la distribución de la frecuencia y la severidad por separado.
3. Se estudiará el comportamiento de las variables respuesta en las distintas categorías representadas.
4. Se estudiará la estructura de correlación de las variables numéricas con el objetivo de identificar cuáles tienen correlación con la variable respuesta y mejorar la futura selección de variables y disminuir la posibilidad de tener variables que sean combinaciones lineales de otras o, en último caso, identificar variables aptas para realizar análisis de componentes principales.

3.3. Modelación

Ya estudiados los patrones existentes, el siguiente objetivo es realizar la etapa de modelación, la cual variará según la variable respuesta que estemos estudiando, el número de siniestros y el valor reembolsado, puesto que cada variable tendrá patrones y relaciones distintas.

Pre-Procesamiento de Datos: Inicialmente, se definirán las variables predictoras de los modelos a estudiar. Luego, la base se dividirá en dos subconjuntos: el primero, que contará con la base completa menos los últimos cuatro meses del último año, se utilizará para entrenar los modelos. El segundo set, que cuenta con el resto de observaciones, se utilizará para validar el poder predictivo o de clasificación de los modelos entrenados anteriormente.

3.3.1. Estrategia de Modelación

Durante esta etapa se plantearán los modelos y la estrategia de modelación que se utilizará para abordar el estudio de las variables respuesta basado en el análisis descriptivo que se realizará:

Número de Siniestros Corresponde a una variable de conteo que representa el número de veces que se fue al médico en X agrupación, por lo cual se espera utilizar modelos con la capacidad de tratar con este tipo de datos, ya que están diseñados para esa distribución en específico o que la función de coste esté diseñada para este problema.

Reembolso Aplicado Variable de carácter continuo, por lo que se utilizarán modelos que puedan tratar con una respuesta continua, y en caso de ser necesario utilizar algún modelo que trate con una distribución específica.

La estrategia de modelación para los modelos asociados a las variables respuesta consistirá en que para cada modelo se empezará utilizando un modelo con todas las variables de estudio y luego se irán probando combinaciones de predictoras de menor dimensión con el objetivo de encontrar la combinación óptima.

3.3.2. Etapa de Evaluación del Modelo

En esta etapa se explicarán las métricas utilizadas para evaluar los modelos estudiados y gráficos tanto de residuos como de comparación para estudiar la efectividad del modelo, con el objetivo de decidir la mejor opción. Las métricas que se utilizarán son las siguientes:

- **Error Cuadrático Medio (MSE)**: Corresponde al promedio de la suma cuadrática de los errores del modelo, introducido para penalizar los modelos que presenten diferencias de mayor magnitud.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Error Medio Absoluto (MAE)**: Corresponde al promedio de la diferencia absoluta entre los valores reales y los predichos por el modelo.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Error Porcentual Absoluto Medio (MAPE)**: Indica en promedio, cuánta diferencia porcentual hay entre los valores reales y predichos.

$$\text{MAPE} = \frac{100\,\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **Criterio de Información de Akaike (AIC)**: Ofrece una estimación relativa de la información perdida, en el sentido de que evalúa cuánto del contenido informativo de los datos se pierde al realizar

un ajuste de un modelo.

$$\mathrm{AIC} = -2\ln\left(\hat{L}\right) + 2k \tag{1}$$

Capítulo 4

Resultados Obtenidos

En este capítulo se presentarán y discutirán los resultados obtenidos en cada etapa planteada en la sección 3.

4.1. Resultados Desarrollo Base de Datos

Durante el desarrollo de la primera etapa, se generaron las siguientes bases de datos:

4.1.1. Recolección y Unión de Datos

- **Sábana de Gastos**: Se tomó información de la sábana de gastos de la compañía que cumplieran los siguientes requisitos:

  1. Que tuviera las variables mencionadas en 3.1.
  2. Que fuese de un siniestro cuyo reembolso fue aceptado y pagado por la compañía.
  3. Que perteneciera a una póliza colectiva, puesto que la sábana contenía información de ambas.
  4. Que los siniestros pertenecieran fueran pagados entre febrero-2020 y Agosto-2024, pudiendo ampliarse en futuros estudios.
  5. Se filtran los códigos de productos asociados a consultas médicas ambulatorias.

  La información antes de ser procesada tiene un total de 251.480 observaciones.

- **Archivo de Cobranzas**: La base de cobranza se compone de la unión de múltiples archivos que entregan información de los cobros hacia la compañía mensual para cada póliza, pero lo que interesa de esos archivos es que se puede desprender el género de los asegurados, para lo cual se siguió el siguiente procedimiento para poder procesarlos:

  1. Se unifica el formato de los archivos.
  2. En cada archivo se extraen las tablas con la información requerida.
  3. Se unen en un archivo único todas las tablas extraídas.
  4. Se filtra la información requerida y se eliminan duplicados.

  Luego de procesarlas, el archivo final se compone de 1.116.015 observaciones.

- **Evolución de Asegurados**: Para la sábana de la compañía se obtuvo en un archivo ya unificado, como se mencionó en 3.1.1, de la cual se tiene información histórica desde el año 2016 hasta la actualidad, de los cuales se filtraron los periodos febrero-2020 y Agosto-2024.
- **Información Geográfica**: Como se menciona en 3.1, la información se obtiene directamente del Servicio de Impuestos Internos (SII), por lo cual la información viene lista para su uso.

4.1.2. Limpieza de Datos

Durante esta sección se detallarán los mecanismos usados para dejar cada base recopilada de tal forma que se pueda dividir, según el tipo de variable en estudio:

- Variables Cualitativas: Inicialmente se estudia la cantidad de categorías redundantes y con poca presencia de observaciones (<500) y se recategorizan las categorías.
- Variables Cuantitativas: Se estudió cada variable individualmente y se analiza la existencia de datos que hayan sido mal escritos o incongruentes con el conjunto en total. En casos específicos, se reestructuró el formato para uso futuro:
    1. La variable asociada a la fecha venía en muchas estructuras distintas, se tenía tanto en formato texto como numérico y se decidió dejar en formato $mm - yyyy$ para un manejo más cómodo.
    2. La variable asociada al rut del Titular, se puede encontrar en el siguiente formato "10.000.000−1", al que se le eliminan los puntos, el guion y el dígito identificador. En caso de estar en este formato "10000000 − 1" se elimina solo el guion y el dígito que le sigue. Y se transforma la variable a una de tipo carácter.

4.1.3. Resultados del cruce de Base de datos

En la siguiente sección se discutirá la pérdida de observaciones que se logró observar al momento de realizar la metodología propuesta en 3.1.3:

- **Cruce Base de SII**: Al momento de realizar el cruce, se obtuvo una pérdida de 5.671 observaciones, asociadas a 5.395 prestadores de salud, los cuales provenían de prestadores particulares (médicos, personas particulares, etc...) cuyo rut no estaba en la base de datos.
- **Cruce con Cobranza**: Al momento de realizar el cruce, se obtuvo una pérdida de 43.671 observaciones, producto de que los datos recolectados no lograron capturar el total de la población. Este resultado se puede mejorar para estudios futuros, pero la información que entrega es suficiente para poder proseguir con el estudio.

4.1.4. Reducción de la Información

En esta etapa se discutirán los primeros resultados obtenidos de la base ya finalizada. Esta comprendió finalmente un total de 214.859 observaciones sin reducir y 57.981 observaciones reducidas. Luego, se analizó la existencia de valores incoherentes, es decir, valores extremadamente altos, en las cuales se encontró que en el reembolso aplicado existían observaciones mal clasificadas, puesto que en el diagnóstico estaban como exámenes médicos. Estas observaciones constaban de menos de 100 y fueron eliminadas de la base de estudio.

4.2. Análisis Exploratorio

El segundo objetivo a cumplir comprende el análisis de la base de datos propuesta, buscando encontrar asociaciones e información preliminar que nos permita encaminar el uso de distintas técnicas estadísticas para problemas de regresión o clasificación.

4.2.1. Descripción Base de Datos

Tabla 1: Descripción Variables Cualitativas

Variables Cualitativas		
Variable	**Descripción**	**Categorías**
Servicio	Especialidad de la Consulta	-General -Broncopulmonar -Cardiología -Dermatología - Endocrinología -Gastroenterología -Ginecología - Medicina Familiar -Medicina Interna -Otras -Neurología -Oftalmología -Otorrinolaringología - Pediatría -Psicología - Respiratorias -Reumatología - Traumatología -Urología
Género	Género del Asegurado	-Femenino -Masculino -Desconocido
Relacion	Relacion con el Titular	-Titular -Conyuge -Hijo
Año	Año donde se pagó el siniestro	2020:2024
Mes	Mes donde se pagó el siniestro	Enero:Diciembre
Región	Región donde reside el Prestador	-Metropolitana -Otras Regiones
Comuna	Comuna donde reside el Prestador	-Metropolitana -Otras Regiones
Primera Capa IAgrupado	Categoriza el origen del seguro de salud obligatorio del asegurado	-Fonasa -Isapre -Otros
Red de Prestadores	Nombre del Prestador de salud	-Empresas Red Salud S.A. -Otros -Empresas Banmédica -Condes -ACHS Salud -Bupa Chile -Andes Salud -Red Interclínica -UC -Grupo Alemana -Meds

Tabla 2: Descripción Variables Cuantitativas

Variables numéricas	
Variable	**Descripción**
NSiniestros	Cantidad de siniestros por agrupación
ValorPagoUF	Reembolso realizado por siniestro en UF
ValorPrestacion	Valor original del servicio en UF
BonIsapreUF	Valor Bonificado por Fonasa o Isapre en UF
DeducibleUF	Valor del deducible pagado por el asegurado en UF

4.2.2. Distribución Variables Respuesta

En esta sección se mostrará la distribución de las variables objetivo a estudiar, con el objetivo de caracterizar la variable y la forma que tiene su distribución, y encontrar alguna transformación que sea más cómoda de manejar.

4.2.2.1. Número de Siniestros

Corresponde a la variable a modelar si se requiere estudiar la frecuencia, debido a que corresponde al número de siniestros realizados y pagados en alguna agrupación. Esta se distribuye de la siguiente manera:

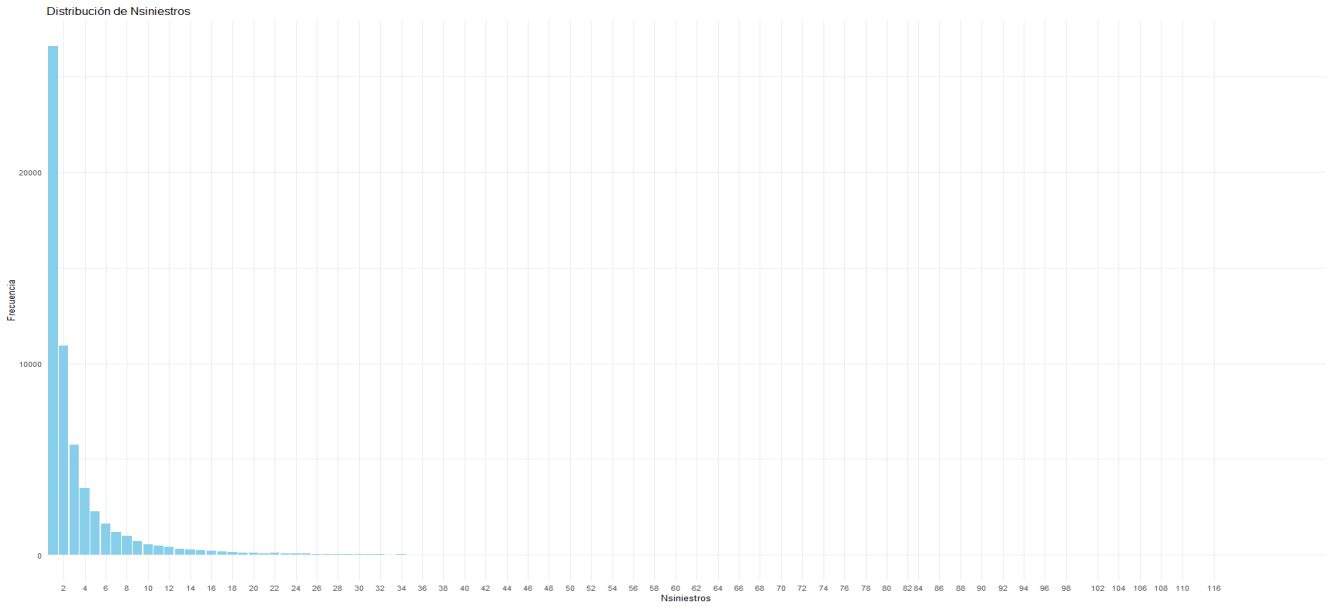

Figura 2: Distribución del número de Siniestros

Como se puede observar, tiene un comportamiento como una variable de conteo, pero con una gran cantidad de datos outliers, que indican que existen agrupaciones que generan una gran cantidad de siniestros. Por lo tanto, pesan mucho en la base de datos, por lo que la futura modelación debe tratar con cuidado este problema y abordarlo con funciones de costo que no sean tan sensibles a los outliers.

4.2.2.2. Reembolso Aplicado

Variable que corresponde al reembolso que recibe el titular del seguro. Esta corresponde a una variable continua que se distribuye de la siguiente manera:

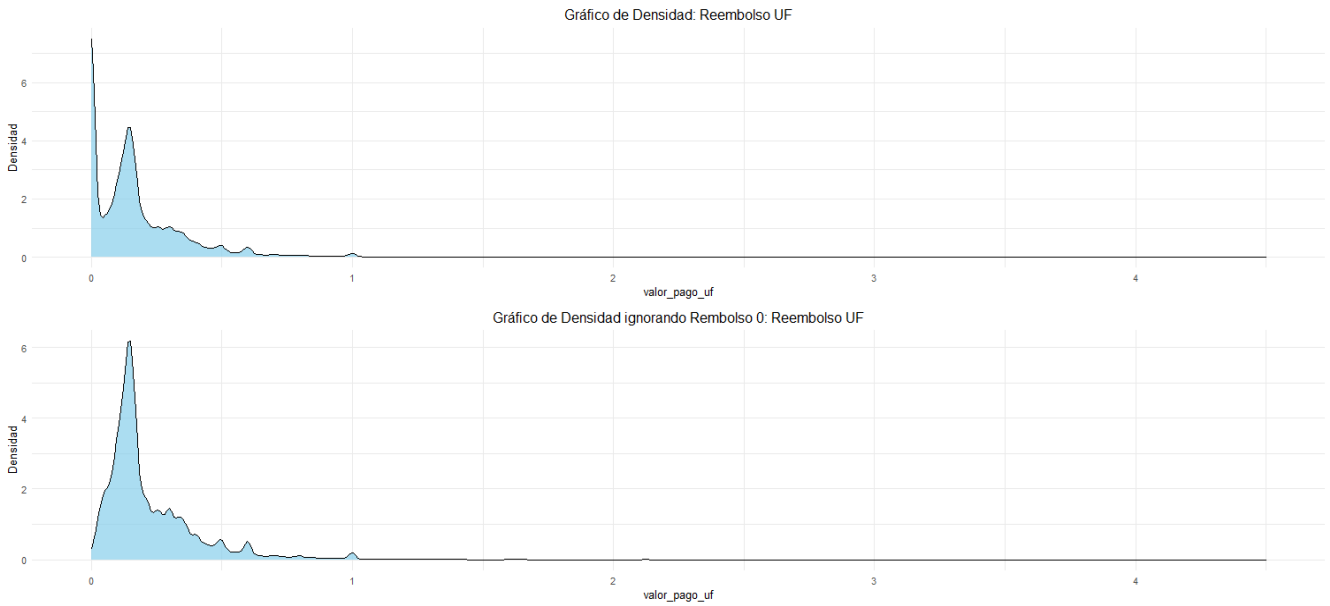

Figura 3: Distribución del Reembolso Aplicado

Tabla 3: Estadísticas descriptivas de Reembolso Aplicado

Vector	Media	Mediana	Desviación Est.	Mínimo	Máximo	Cuartil 25 %	Cuartil 50 %	Cuartil 75 %	Coef. de Var.	Curtosis
Reembolso	0.1682	0.1351	0.1923	0.0000	4.5003	0.0000	0.1351	0.2231	114.3230	19.1525
>0	0.2266	0.1593	0.1913	0.0001	4.5003	0.1169	0.1593	0.2904	84.4413	21.9176

Como se puede observar, corresponde a una distribución con una inflación de ceros, debido a distintos factores, existen situaciones en las que el titular no recibe reembolso, por ejemplo, que la primera capa cubre el total, por topes de la póliza, entre otros factores. Esto, al utilizar modelos clásicos, corresponde a un problema, ya que la gran cantidad de ceros puede confundir las relaciones y cargarlas a valores cercanos al cero. Para evitar este problema, se filtrarán los casos en los que hubo un reembolso mayor que cero, en el cual se observa una forma similar a una distribución normal, pero con una cola superior mucho mayor a la mediana, producto de la cantidad de outliers. Para solventar este problema, se utilizó una transformación logarítmica (figura 4 y tabla 4).

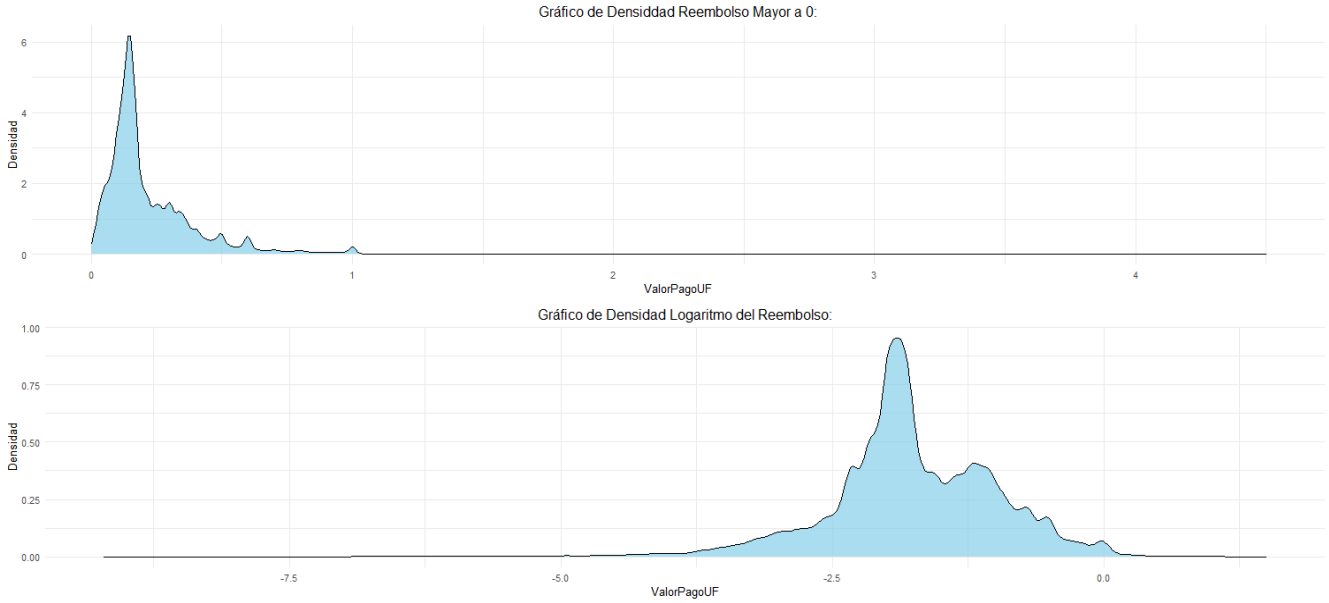

Figura 4: Distribución del Reembolso Aplicado log()

Tabla 4: Estadísticas descriptivas de Reembolso Aplicado log()

Vector	Media	Mediana	Desviación Est.	Mínimo	Máximo	Cuartil 25 %	Cuartil 50 %	Cuartil 75 %	Coef. de Var.	Curtosis
Reembolso	0.1682	0.1351	0.1923	0.0000	4.5003	0.0000	0.1351	0.2231	114.3230	19.1525
log()	-1.7642	-1.8370	0.7766	-9.2103	1.5041	-2.1464	-1.8370	-1.2365	-44.0216	5.8322

Se observa que la transformación, aunque es capaz de penalizar los outliers superiores y controlar mejor la curtosis, termina creando una cola inferior excesivamente pesada, y se observa que la distribución cambia a ser multimodal, lo cual es menos cómodo para modelar. Dado lo anterior, se utilizará la variable sin transformar, y la estrategia para abordar esto será utilizar un modelo de regresión múltiple como base del estudio y modelos de machine learning para abordar la modelación.

4.2.3. Distribución Variables Cualitativas

Durante esta sección se introducirán las posibles covariables de carácter cualitativo del estudio, en las que se presentará su definición, su distribuyen en la base de datos y su relación con las variables a explicar.

4.2.3.1. Relación con el Titular

Indica el tipo de relación familiar que tiene el beneficiario con el titular del seguro, lo cual es importante debido a que no solo el titular hace uso de este, por consiguiente, sus cargas deben ser tenidas en cuenta al momento de determinar la prima. El comportamiento de la variable se observa en la figura 5:

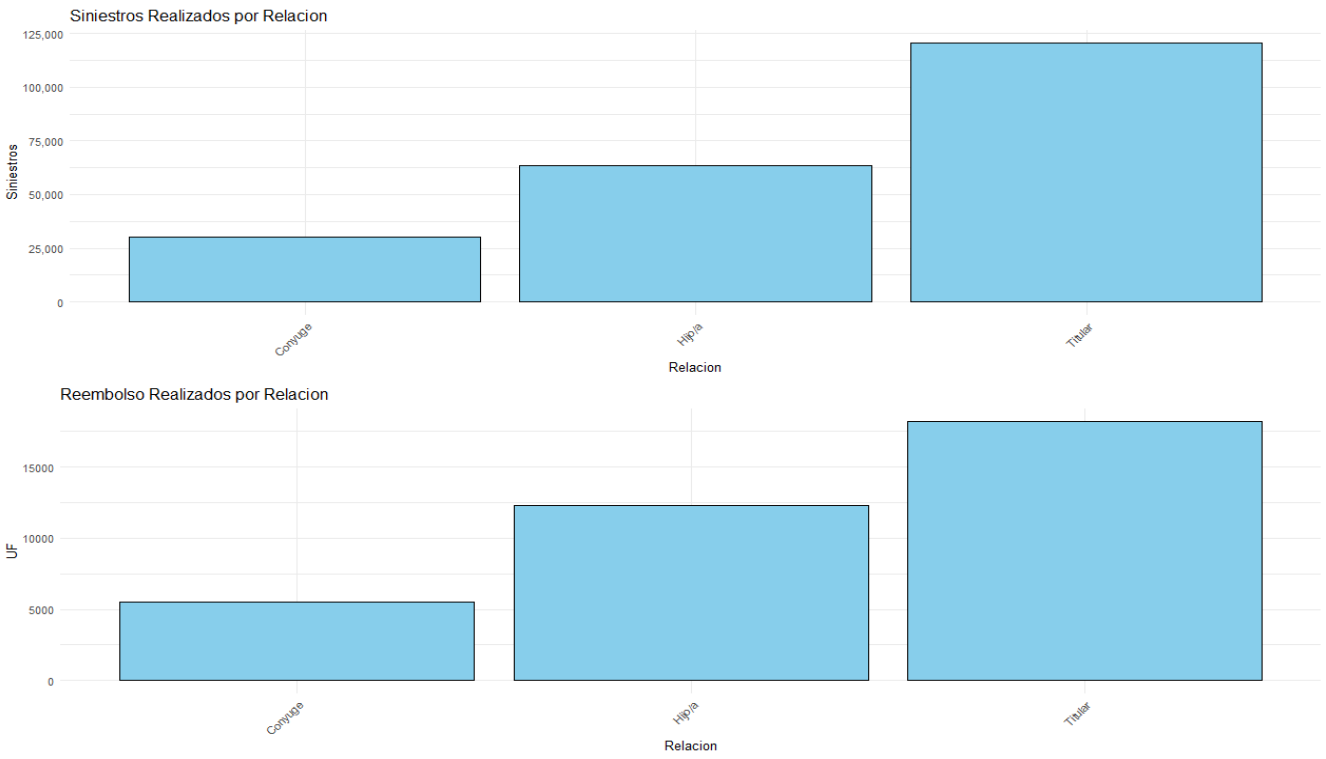

Figura 5: Distribución por Categoría de Relación

Donde se puede observar que la proporción de observaciones indica que la mayoría de siniestros están asociados a los titulares e hijos, y que la cantidad de reembolso no varía a simple vista. Sin embargo, si se analizan las categorías individualmente, podemos encontrar que:

Número de Siniestros Con el objetivo de estudiar si existen cambios significativos entre el tipo de beneficiario del seguro, se mostrará el número de siniestros por categoría (figura 6 y tabla 5):

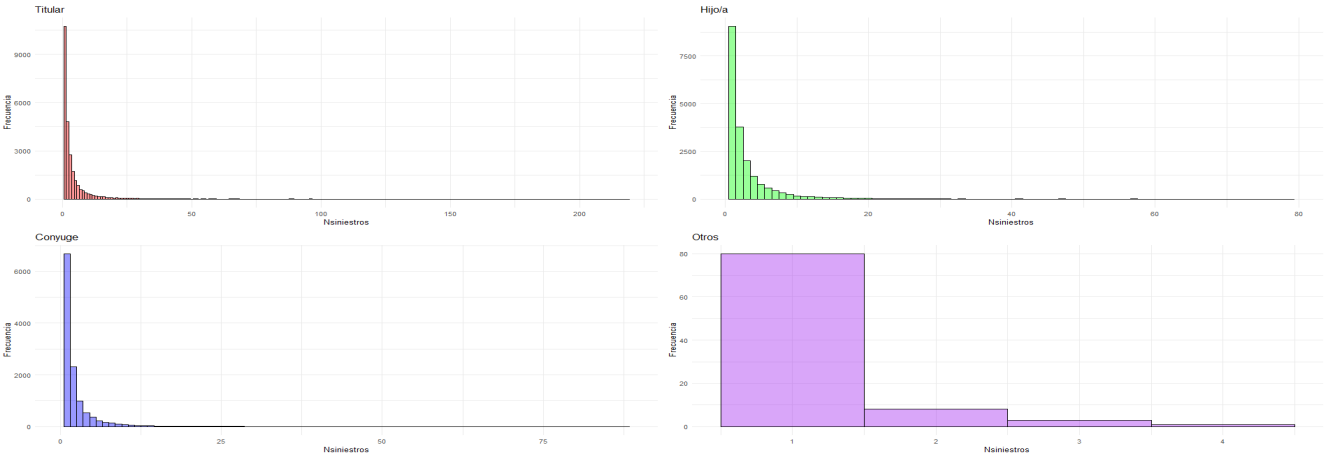

Figura 6: Distribución número de siniestros por Categoría de Relación

Tabla 5: Resumen Descriptivo para Siniestros por Categoría de Relación

Relación	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Cónyuge	11922	1	1	1	2.54	16.99	2	88
Hijo/a	19619	1	1	2	3.24	22.32	3	79
Titular	26358	1	1	2	4.57	77.2	4	219

Tabla 6: Resultado de Kruskal-Wallis de Siniestros por Categoría de Relación

Kruskal-Wallis	Kruskal-Wallis: $\chi^2 = 1176.27, df = 2, p =< 2.22e-16$

Se observa que las categorías se comportan de manera similar, pero la categoría hijo muestra una caída más lenta. Además, se puede observar que la observación con mayor frecuencia la tiene el titular, y que, en términos de medias, los grupos mantienen una media con valores dentro de un rango apropiado. Sin embargo, se realizará un test de Kruskal-Wallis para verificar si se cumple H_0, es decir, que la distribución de los grupos es igual. Donde podemos observar(tabla 6) que, bajo una insignificancia del 0.05, al menos un grupo posee una distribución distinta al resto y, por lo tanto, el número de siniestros sí varía dependiendo de su relación con el titular.

Reembolso Aplicado : Ahora, en relación con el reembolso aplicado (figura 7 y tabla 7):

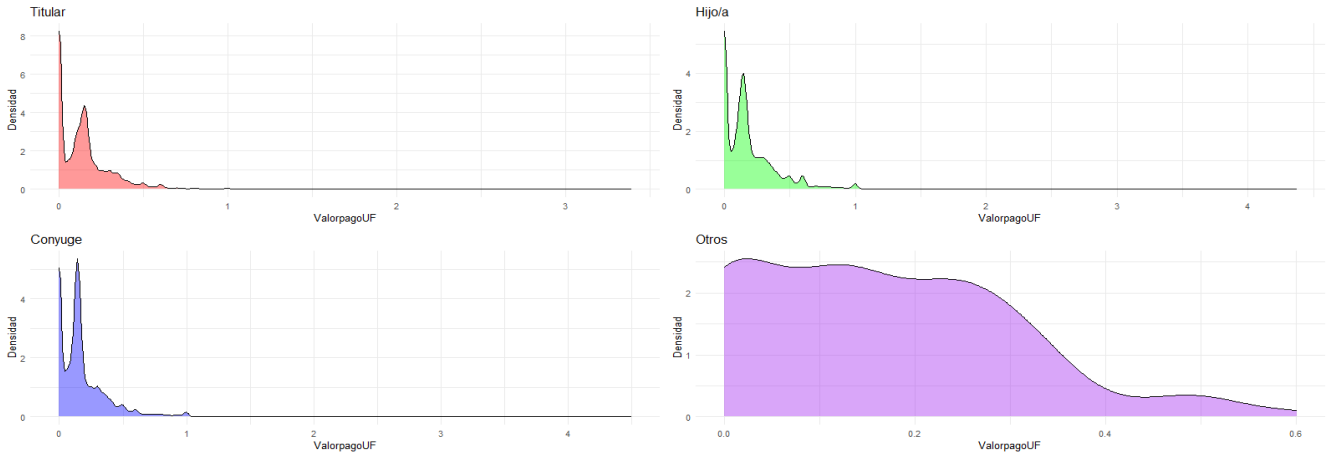

Figura 7: Distribución del Reembolso por Categoría de Relación

Tabla 7: Resumen Descriptivo de Reembolso Aplicado por Categoría de Relación

Relación	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Cónyuge	30233	0	0.05	0.14	0.18	0.04	0.24	4.5
Hijo/a	63603	0	0.02	0.14	0.19	0.05	0.27	4.37
Titular	120539	0	0	0.12	0.15	0.03	0.2	3.39

Tabla 8: Resultado de Kruskal-Wallis para Reembolso por Categoría de Relación

Kruskal-Wallis	Kruskal-Wallis: $\chi^2 = 1673.52, df = 2, p =< 2.22e-16$

En este caso, se observan diferencias significativas entre las categorías, donde la media más baja la tiene el titular, debido a la gran cantidad de ceros en la distribución. Además, se observa que cónyuge tiene una cantidad de ceros mucho menor que el resto de las categorías. En relación con el test de Kruskal-Wallis, podemos observar que(tabla 8), bajo un nivel de significancia del 0.05, al menos un grupo posee una distribución distinta al resto, y por lo tanto, el valor del reembolso sí varía dependiendo de su relación con el titular.

4.2.3.2. Primera Capa

Variable que describe el origen de la primera capa del beneficiario del seguro, esta comprende tres categorías: Isapre, Fonasa y Otros. La tercera está compuesta por sistemas asociados al ejército e instituciones desconocidas. Su distribución está dada por(figura 8):

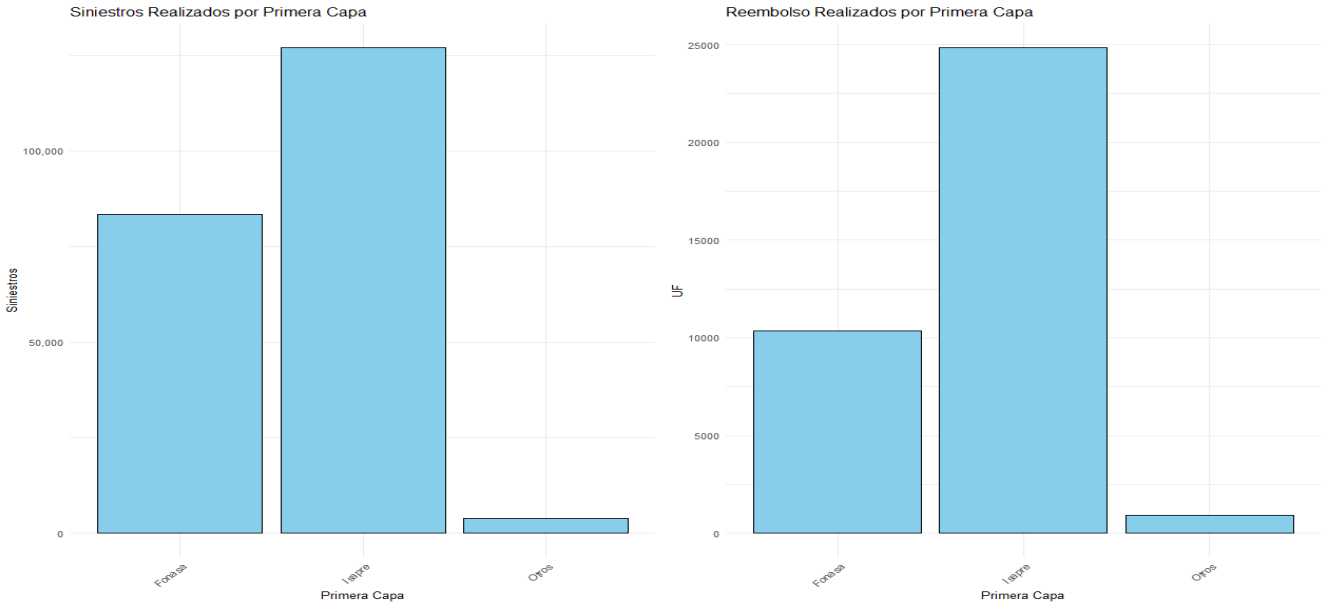

Figura 8: Distribución por Categoría de Primera Capa

Se observa que la distribución es similar, pero que en el reembolso, se aumenta la diferencia entre Isapre y Fonasa, lo cual tiene sentido, debido a que, al ser Isapre un sistema privado, la bonificación que entrega a los beneficiarios es mucho mayor. Además, la gran diferencia en número de siniestros tiene sentido, puesto que, al tener que desembolsar más para poder pagar un servicio, los beneficiarios con Fonasa tienden a ir menos a los sistemas de salud privados.

Número de Siniestros : Se estudiará cómo se comporta el número de siniestros en cada categoría(figura 9 y tabla 9):

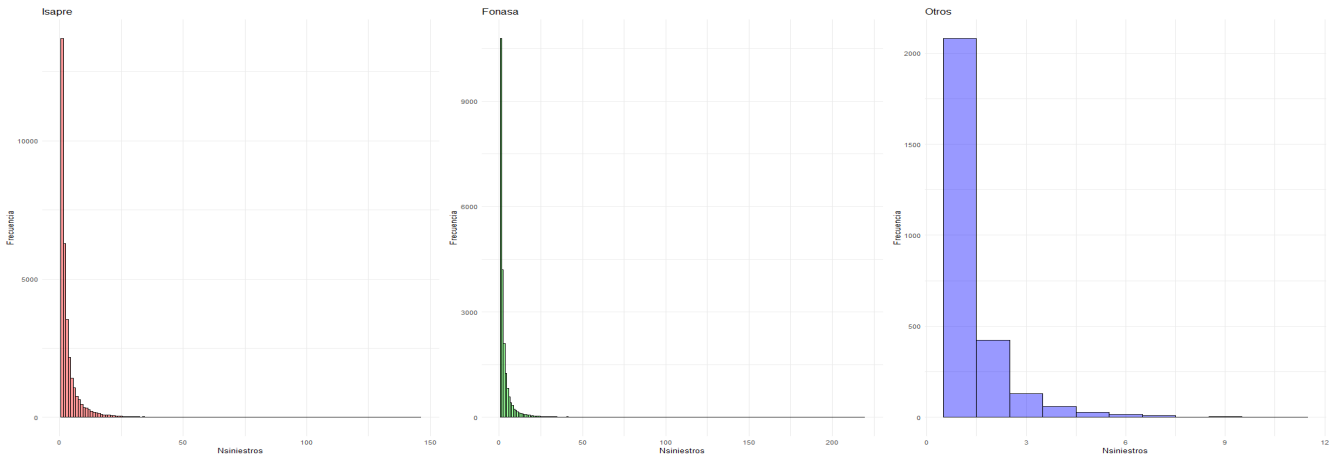

Figura 9: Distribución número de Siniestros por Categoría de Primera Capa

Tabla 9: Resumen Descriptivo de Siniestros por Categoría de Primera Capa

Primera Capa	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Fonasa	22428	1	1	2	3.72	56.5	3	219
Isapre	32811	1	1	2	3.87	43.62	4	146
Otros	2752	1	1	1	1.42	0.92	1	11

Tabla 10: Resultado de Kruskal-Wallis para Siniestros por Categoría de Primera Capa

Kruskal-Wallis	Kruskal-Wallis: $\chi^2 = 1411.94, df = 2, p =< 2.22e-16$

Se observa que el comportamiento es similar tanto en Isapre como en Fonasa, pero con una caída más lenta en Isapre, lo cual se confirma al observar el tercer cuantil de la tabla. Además, contrario a lo esperado, el grupo más grande no fue el que presentó más varianza, sino que fueron los asegurados de Fonasa. Se aplica el test de Kruskal-Wallis con el objetivo de estudiar si la distribución de los tres grupos es la misma, donde, bajo la hipótesis nula planteada en 2.2.1, y segun lo observado(tabla 10) puede concluir que, bajo una significancia del 0.05, el número de siniestros varía dependiendo de la primera capa que presente el titular.

Reembolso Aplicado: En relación con el comportamiento del reembolso aplicado, se observa lo siguiente(figura 10 y tabla 11):

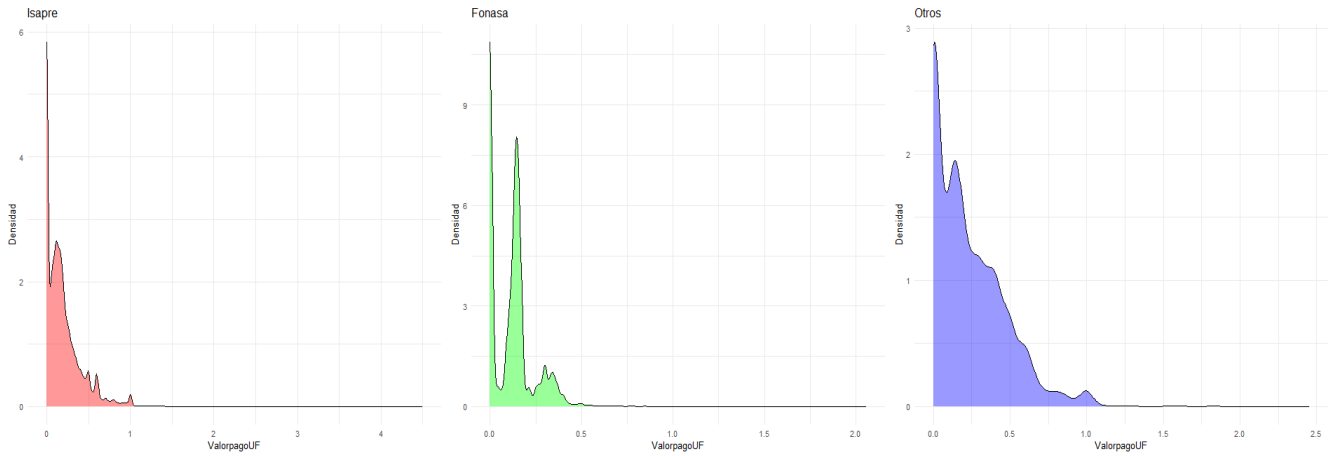

Figura 10: Distribución del Reembolso por Categoría de Primera Capa

Tabla 11: Resumen Descriptivo de Reembolso por Categoría de Primera Capa

Primera Capa	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Fonasa	83525	0	0	0.13	0.12	0.01	0.16	2.06
Isapre	127056	0	0.03	0.14	0.2	0.05	0.27	4.5
Otros	3903	0	0	0.17	0.23	0.06	0.38	2.45

Tabla 12: Resultado de Kruskal-Wallis para Reembolso por Categoría de Primera Capa

Kruskal-Wallis	Kruskal-Wallis: $\chi^2 = 4200.76, df = 2, p =< 2.22e-16$

En este caso, se observan diferencias en el comportamiento, puesto que Fonasa se concentra en valores mucho más bajos de reembolso en comparación con Isapre. Además, Isapre, a diferencia de Fonasa en el número de siniestros, concentra una mayor varianza y valores más altos de reembolso. Para verificar que los grupos se comportan de manera distinta, se aplicará el test de Kruskal-Wallis (tabla 12), que, bajo una significancia del 0.05, indica que el reembolso aplicado sí varía dependiendo de la primera capa que presente el titular.

4.2.3.3. Tipo de Consulta/Servicio

Variable aplicada para definir el tipo de consulta médica de la cual se originó el servicio, puesto que existen las especialidades y el precio varía entre ellas. Esta variable no se aplica directamente en la tarificación, sino que se usa el precio por la agrupación, en este caso, consultas médicas, pero desglosar esta información puede ayudar a entregar valores más precisos. Su distribución está dada por(figura 11) .

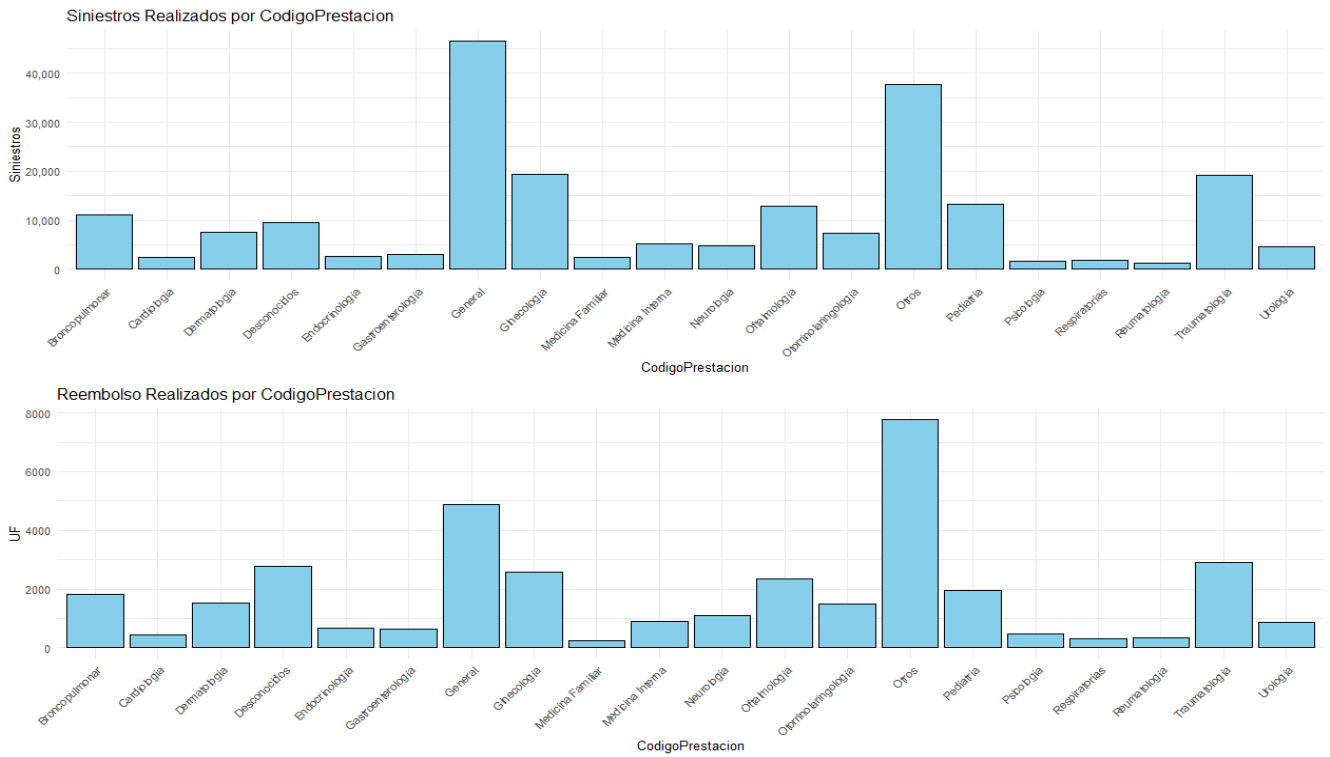

Figura 11: Distribución por Categoría de Tipo de Consulta

Se puede observar que la mayor cantidad de observaciones están relacionadas con consultas generales, traumatismo, ginecología y otros, cuyas categorías se desconocen o son muy pequeñas. En relación al reembolso, se observa que otros es mucho mayor en comparación con otras categorías. A excepción de lo mencionado, no se encuentran diferencias a simple vista.

Número de Siniestros : Al estudiar el comportamiento de las categorías con mayor número de observaciones(figura 12 y tabla 13), se encuentra que:

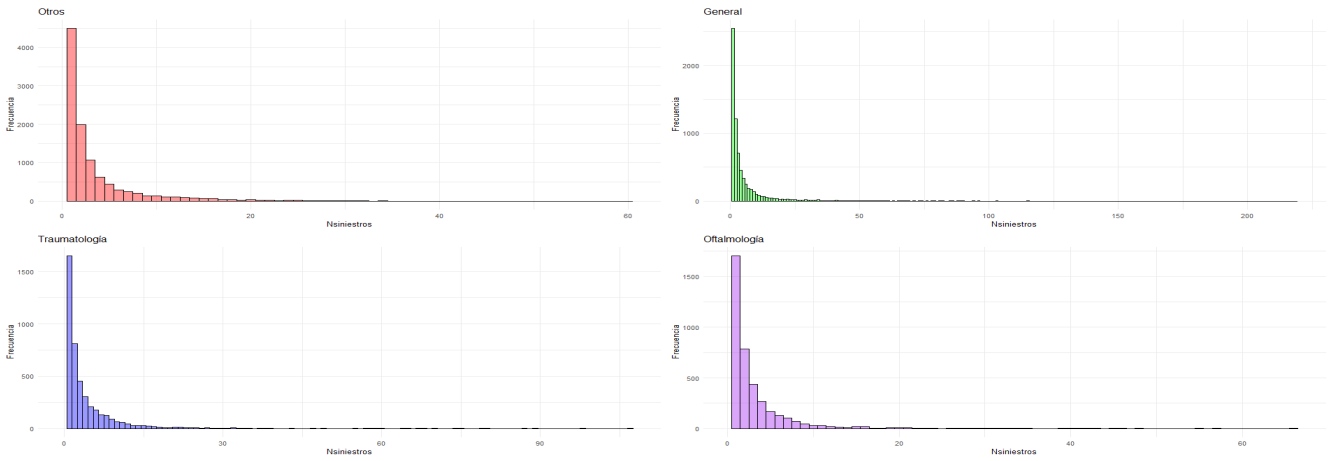

Figura 12: Distribución del número de Siniestros por Categoría de Tipo de Consulta

Tabla 13: Resumen Descriptivo de Siniestros por Categoría de Tipo Consulta

Tipo Consulta	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
General	7097	1	1	2	6.55	175.09	6	219
Oftalmología	3942	1	1	2	3.27	20.61	4	66
Otros	10438	1	1	2	3.61	22.34	4	60
Traumatología	4401	1	1	2	4.34	48.15	5	107

Tabla 14: Resultado de Kruskal-Wallis para Siniestros por Categoría de Tipo Consulta

Kruskal-Wallis	χ^2=278.43, df=3, p=< 2.22e-16

Podemos observar que las categorías con mayor número de siniestros poseen una caída de velocidad similar, puesto que comparten la misma mediana. Además, se puede observar que Otros y Oftalmología se comportan de manera muy similar, aun fuera de la mediana. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla 14), que, bajo una significancia del 0.05, indica que el número de siniestros sí varía dependiendo del tipo de servicio.

Reembolso Aplicado : Al estudiar el comportamiento(figura 13 y tabla 15), se observa lo siguiente:

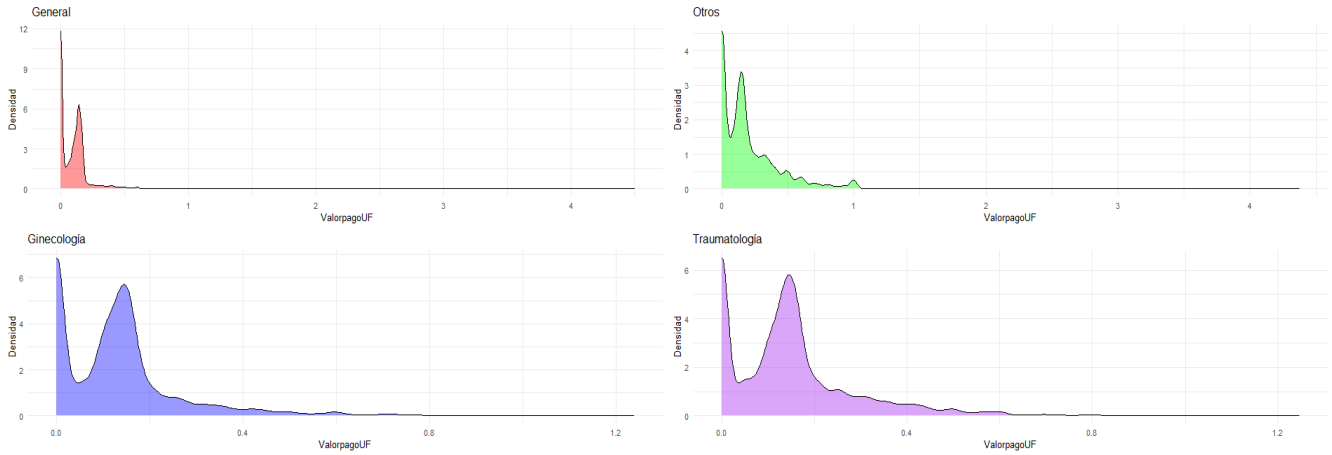

Figura 13: Distribución del Reembolso por Categoría de Tipo de Consulta

Tabla 15: Resumen Descriptivo de Reembolso por Categoría de Tipo Consulta

Tipo Consulta	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
General	46464	0	0	0.1	0.11	0.02	0.14	4.5
Ginecología	19382	0	0	0.12	0.13	0.02	0.16	1.24
Otros	37674	0	0	0.15	0.21	0.06	0.29	4.37
Traumatología	19105	0	0.05	0.14	0.15	0.02	0.19	1.25

Tabla 16: Resultado de Kruskal-Wallis para Reembolso por Categoría de Tipo Consulta

Kruskal-Wallis	χ^2=4948.75, df=3, p=< 2.22e-16

Se observa que las categorías de Traumatología y Ginecología se comportan de manera similar, lo que puede indicar que, en grupos más pequeños, el efecto de los servicios no es tan grande. A diferencia de Otros, donde la diferencia es más marcada. Para verificar que los grupos se comportan de manera distinta, se aplicará el test de Kruskal-Wallis (tabla 16), que, bajo una significancia del 0.05, indica que el reembolso aplicado sí varía dependiendo de la del tipo de consulta.

4.2.3.4. Por Prestador de Salud

Variable introducida para evaluar el peso de los distintos prestadores al estudiar las variables respuesta. Es de interés debido a los convenios existentes entre clínicas y para conocer si impactan en las clínicas más frecuentadas y potenciar las que no lo están. En el caso del reembolso, se estudian las clínicas más caras y si los incentivos a las clínicas asociadas con Banmédica surten efecto. Estas se agruparon según su mayor interés y se distribuyen de la siguiente manera(figura 14):

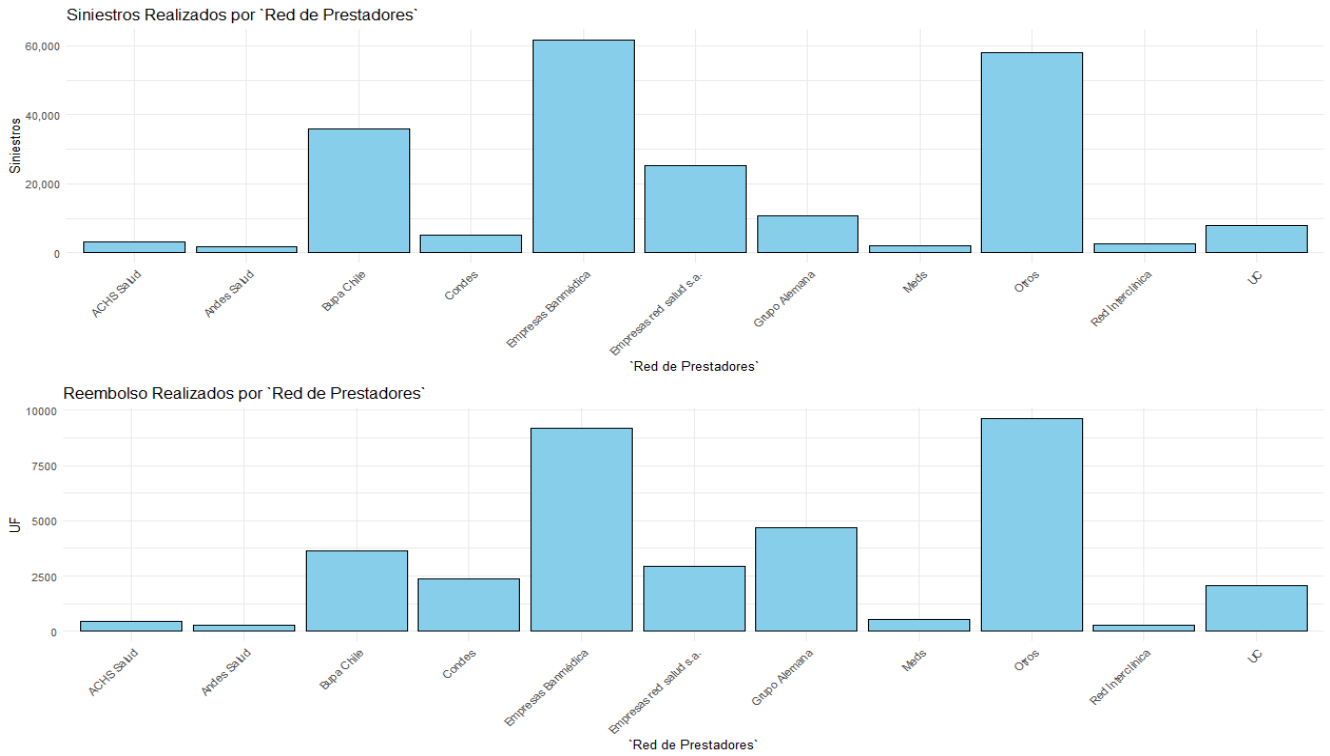

Figura 14: Distribución por Categoría Prestador de Salud

Debido a la gran cantidad de prestadores en el país, se agruparon aquellos con mayor tasa de siniestros, y los más pequeños, que no tienen más de mil siniestros en el período de estudio, fueron trasladados a la categoría Otros. Se observa que la mayor parte de la cartera está compuesta por clínicas y prestadores asociados a Banmédica, lo cual es positivo. Además, se observa que es la que contiene una mayor tasa de reembolso.

Número de Siniestros : Al estudiar el comportamiento de las categorías con mayor número de observaciones(figura 15 y tabla 17), se encuentra que:

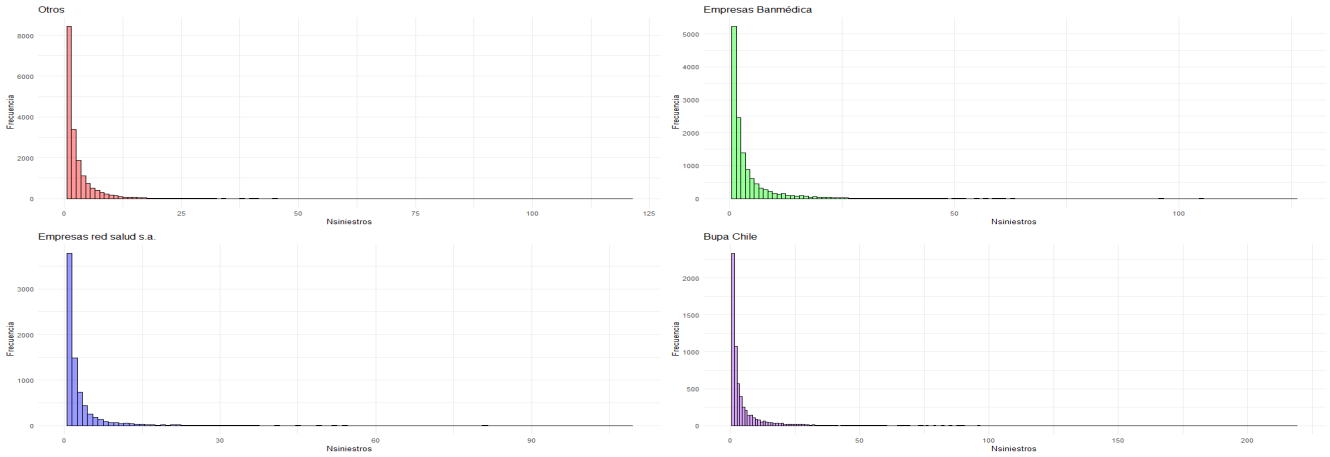

Figura 15: Distribución del número de Siniestros por Categoría de Prestador

Tabla 17: Resumen Descriptivo de Siniestros por Categoría de Prestador

Prestador	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Bupa Chile	6080	1	1	2	5.92	149.87	5	219
Empresas Banmédica	13321	1	1	2	4.64	66.59	5	126
Empresas red salud s.a.	7688	1	1	2	3.28	32.81	3	109
Otros	18097	1	1	2	3.21	25.59	3	121

Tabla 18: Resultado de Kruskal-Wallis para Siniestros por Categoría de Prestador

Kruskal-Wallis	χ^2=610.24, df=3, p=< 2.22e-16

Se puede observar que los principales prestadores tienen un comportamiento similar, destacando a Red Salud y Otros, que comparten un comportamiento más similar entre estas dos categorías. Además, cabe destacar que la clínica Bupa es la más frecuentada por los asegurados. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla 18), que, bajo una significancia del 0.05, indica que el número de siniestros sí varía dependiendo del tipo de prestador.

Reembolso Aplicado: En relación con el comportamiento del reembolso aplicado(figura 16 y tabla 19), se observa lo siguiente:

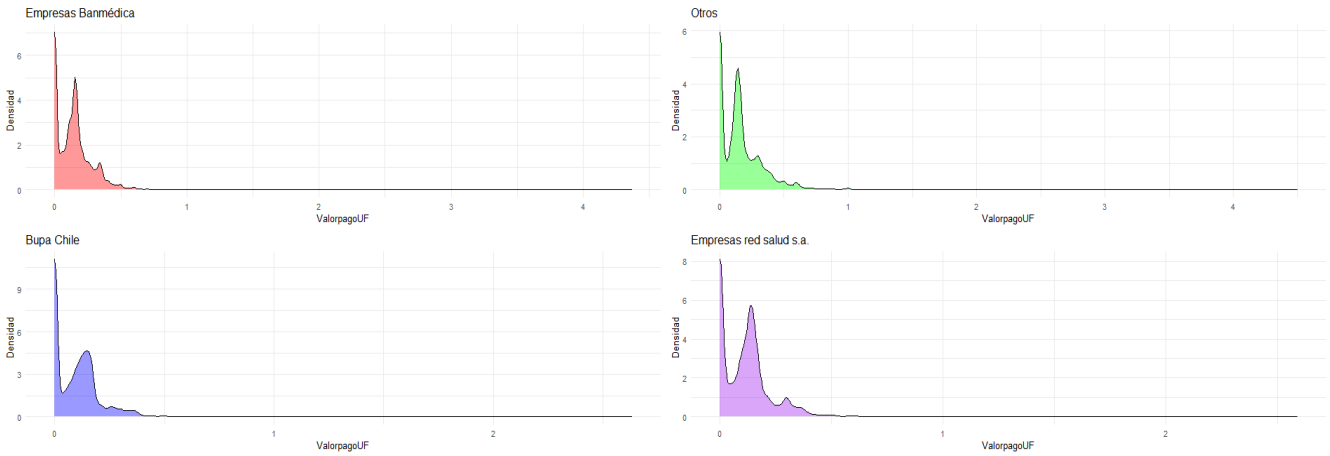

Figura 16: Distribución del Reembolso por Categoría de Prestador

Tabla 19: Resumen Descriptivo de Reembolso por Categoría de Prestador

Prestador	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Bupa Chile	35981	0	0	0.1	0.1	0.01	0.15	2.63
Empresas Banmédica	61773	0	0.02	0.14	0.15	0.02	0.2	4.37
Empresas red salud s.a.	25231	0	0	0.11	0.12	0.01	0.16	2.59
Otros	58095	0	0	0.14	0.17	0.03	0.24	4.5

Tabla 20: Resultado de Kruskal-Wallis para Reembolso por Categoría de Prestador

Kruskal-Wallis	χ^2=4235.59, df=3, p=< 2.22e-16

En este caso, se observa una diferencia solo en la clínica Bupa en comparación con el resto de las categorías, donde tienen una mayor concentración en cero y en valores entre 0 y 0.5 UF. A diferencia de los otros prestadores, no hay un reembolso que predomine más. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla 20), que, bajo una significancia del 0.05, indica que el reembolso aplicado sí varía dependiendo del tipo de prestador.

4.2.3.5. Región del Prestador

La variable que indica la región de origen del prestador, y por ende del origen del siniestro, entrega información de interés, puesto que existen diferencias de precio respecto a Santiago. Esta se distribuye de la siguiente forma (figura 17):

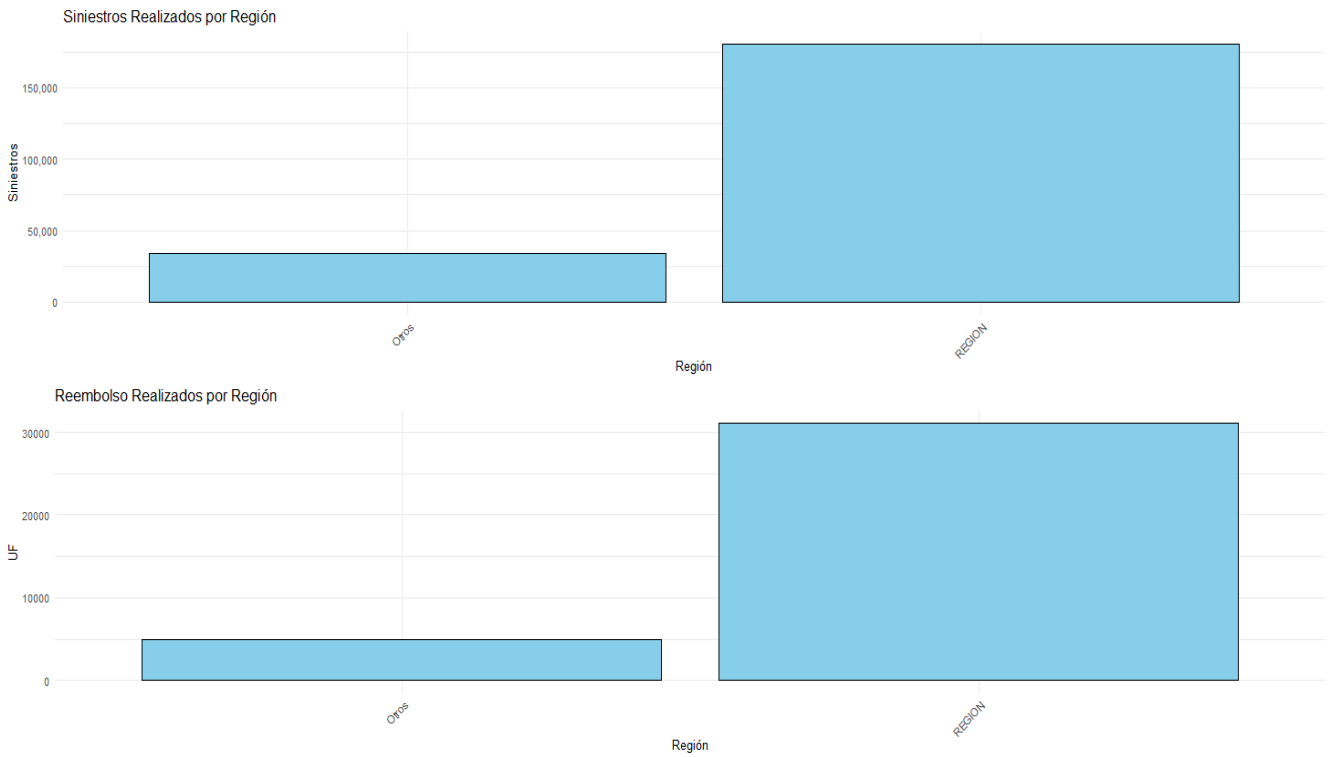

Figura 17: Distribución por Categoría de Región

Donde se observa que la proporción es similar en ambos casos pero se ve mas marcada la diferencia de forma monetaria, es decir, los siniestros que están en región reciben en general menos reembolso.

Número de Siniestros : Al estudiar el comportamiento de las categorías (figura 18 y tabla 21), se encuentra lo siguiente:

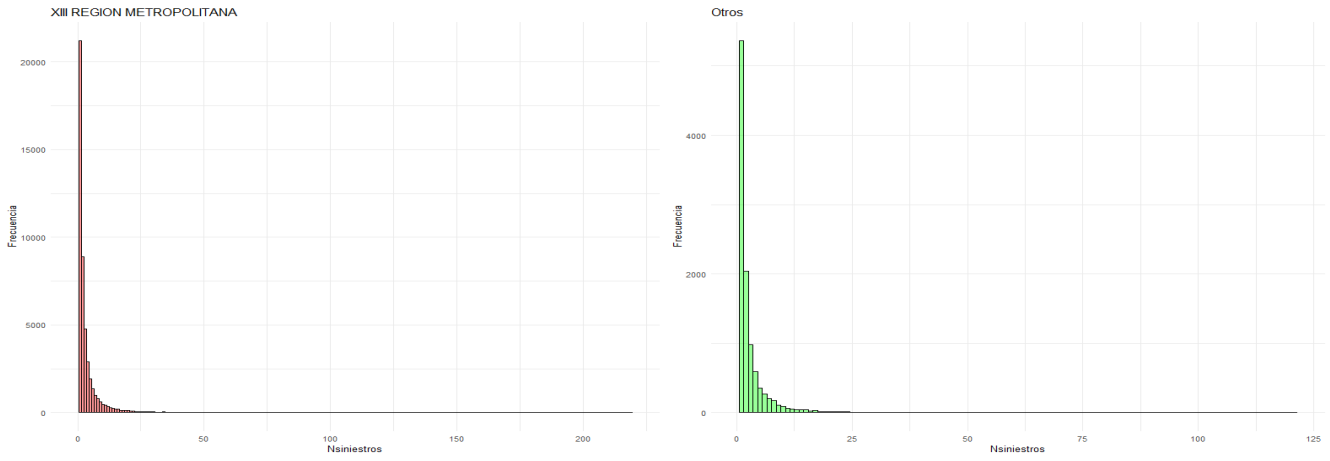

Figura 18: Distribución del número de Siniestros por Categoría de Region

Tabla 21: Resumen Descriptivo de Siniestros por Categoría de Región

Región	n	mínimo	Q1	mediana	media	varianza	Q3	máximo
Otros	10924	1	1	1	3.14	34.58	3	118
Metropolitana	48645	1	1	2	3.71	48.16	3	224

Tabla 22: Resultado de Kruskal-Wallis para Siniestros por Categoría de Región

Kruskal-Wallis	χ^2=128.82, df=1, p=< 2.22e-16

Donde se observa que a pesar de la diferencia de observaciones el comportamiento es similar compartiendo la media y con una varianza similar. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla 22), que, bajo una significancia del 0.05, indica que el número de siniestros sí varía dependiendo de la comuna.

Reembolso Aplicado: En relación con el comportamiento del reembolso aplicado (figura 19 y tabla 23), se observa lo siguiente:

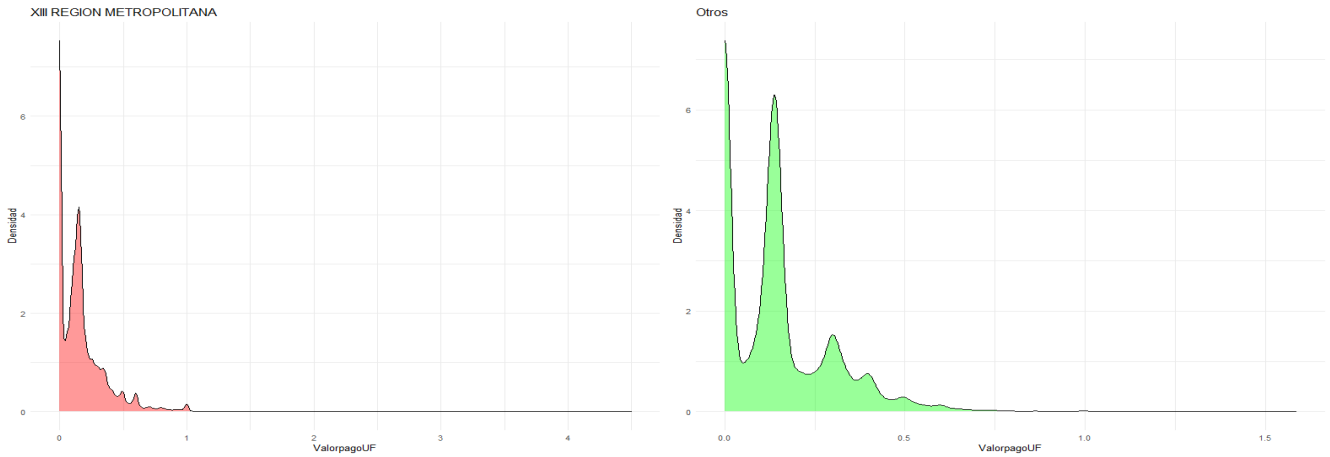

Figura 19: Distribución del número del Reembolso por Categoría de Región

Tabla 23: Resumen Descriptivo de Reembolso por Categoría de Región

Región	n	mínimo	Q1	mediana	media	varianza	Q3	máximo
Otros	34253	0	0	0.13	0.14	0.02	0.2	1.59
Metropolitana	180231	0	0	0.14	0.17	0.04	0.23	4.5

Tabla 24: Resultado de Kruskal-Wallis para Reembolso por Categoría de Región

Kruskal-Wallis	χ^2=380.95, df=1, p=< 2.22e-16

En la figura 19 se observa lo esperado, es decir, que el reembolso aplicado en la región metropolitana es mucho mayor que en regiones, ademas de contar con una media mayor, es la que contiene los datos outlier. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla **??**), que, bajo una significancia del 0.05, indica que el reembolso sí varía dependiendo de la región.

4.2.3.6. Comuna del Prestador

Variable que indica la comuna de origen del prestador. Esta entrega información de gran interés, puesto que proporciona información sobre las comunas con mayor frecuencia, y por sí sola, es relevante para la compañía. Esta se distribuye de la siguiente manera(figura 20):

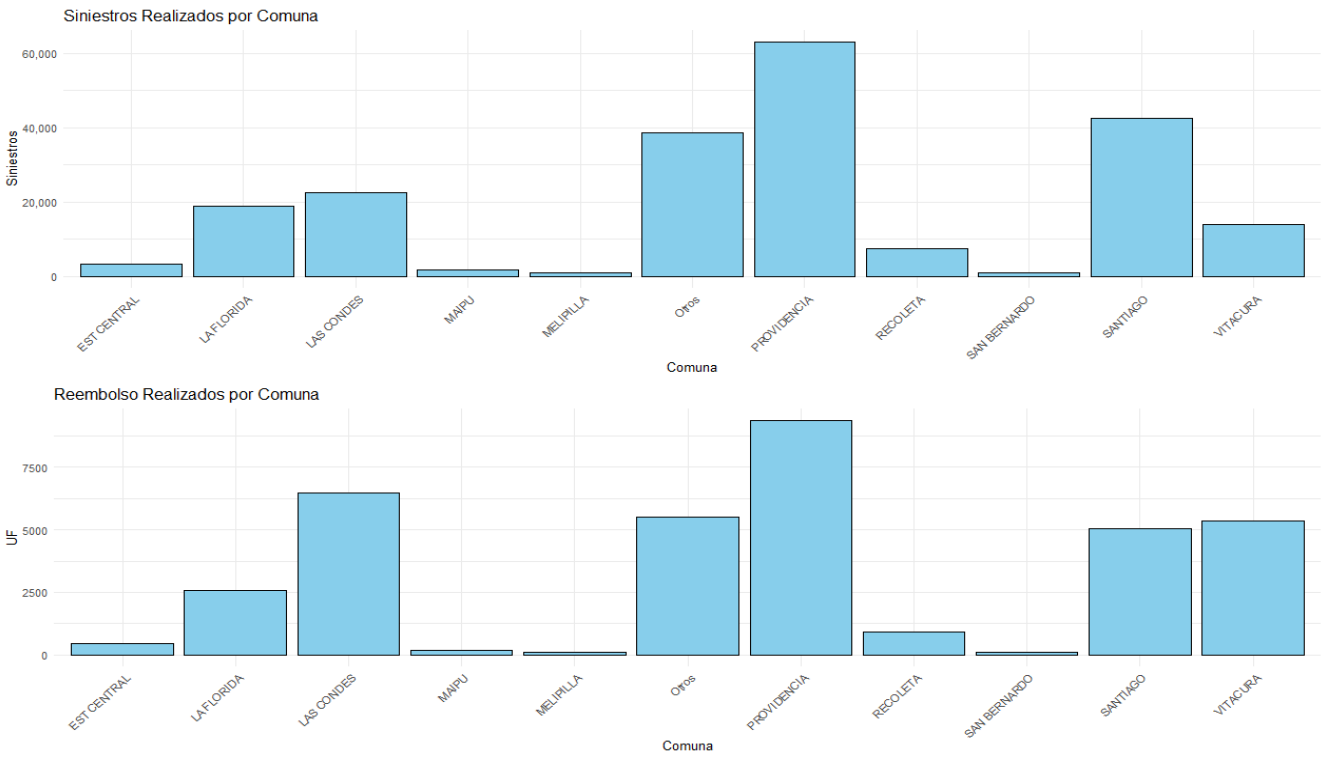

Figura 20: Distribución por Categoría de Comuna

Se destaca que se muestran solo las comunas con un volumen de información suficiente; el resto se agrupa en la categoría Otros. Dado esto, se puede observar lo esperado, que las comunas del sector oriente, con mayor cantidad de población de familias de clase media, son las que tienen mayor número de siniestros y que los centros de salud más concurridos tienen sedes en las comunas más representadas. En cambio, los reembolsos realizados muestran de mejor manera la diferencia en los reembolsos, puesto que comunas con menor cantidad de siniestros, pero que pertenecen a la zona oriente de Santiago, tienen un mayor reembolso.

Número de Siniestros : Al estudiar el comportamiento de las categorías con mayor número de observaciones (figura 21 y tabla 25), se encuentra lo siguiente:

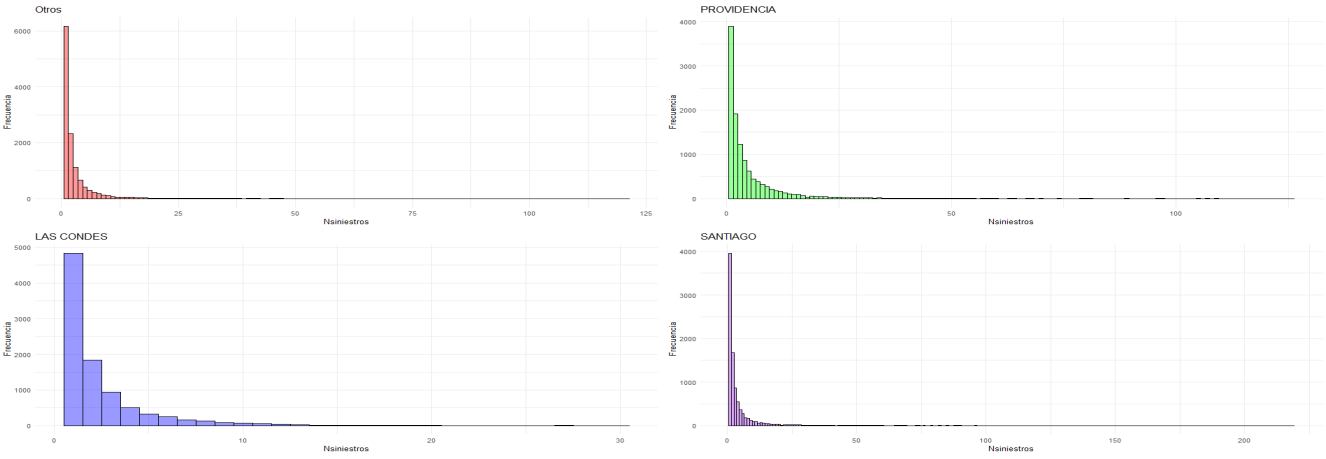

Figura 21: Distribución del número de Siniestros por Categoría de Comuna

Tabla 25: Resumen Descriptivo de Siniestros por Categoría de Comuna

Comuna	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
LAS CONDES	9327	1	1	1	2.43	6.37	3	30
Otros	12194	1	1	1	3.18	32.54	3	121
PROVIDENCIA	11618	1	1	2.5	5.43	78.98	6	126
SANTIAGO	9077	1	1	2	4.69	104.88	4	219

Tabla 26: Resultado de Kruskal-Wallis para Siniestros por Categoría de Comuna

Kruskal-Wallis	χ^2=1668.27, df=3, p=< 2.22e-16

Se observa que, dentro de las comunas con mayor cantidad de siniestros, la diferencia entre la media de la cantidad de siniestros y su variabilidad es muy alta, lo que indica sobredispersión dentro de ellas. La comuna que presenta este problema en menor medida es Las Condes. Además, cabe destacar que la comuna de Providencia tiene una tasa media de siniestros mucho mayor que otras comunas. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla 26), que, bajo una significancia del 0.05, indica que el número de siniestros sí varía dependiendo de la comuna.

Reembolso Aplicado: En relación con el comportamiento del reembolso aplicado (figura 22 y tabla 27), se observa lo siguiente:

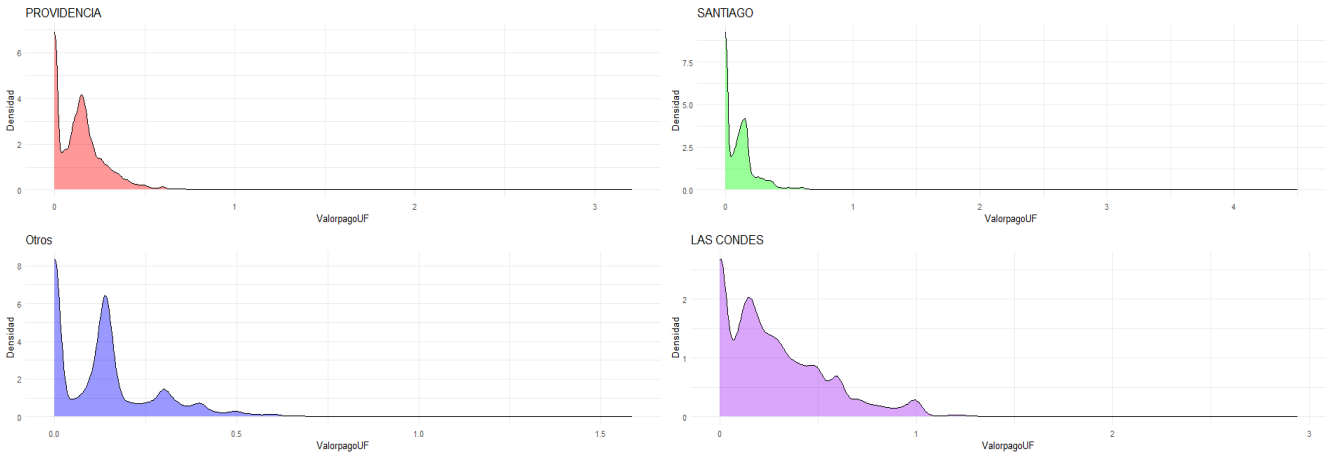

Figura 22: Distribución del Reembolso por Categoría de Comuna

Tabla 27: Resumen Descriptivo de Reembolso por Categoría de Comuna

Comuna	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
LAS CONDES	22646	0	0.08	0.22	0.29	0.07	0.44	2.94
Otros	38740	0	0	0.13	0.14	0.02	0.18	1.59
PROVIDENCIA	63057	0	0.02	0.14	0.15	0.02	0.21	3.21
SANTIAGO	42547	0	0	0.1	0.12	0.02	0.16	4.5

Tabla 28: Resultado de Kruskal-Wallis para Reembolso por Categoría de Comuna

Kruskal-Wallis	χ^2=8300.04, df=3, p=< 2.22e-16

En este caso, se muestran los rubros con una mayor cantidad de reembolso. A diferencia del número de siniestros, se observa que todos tienen una distribución similar, lo cual es un indicio de que se comportan de manera similar entre comunas. Sin embargo, la media entre estas difiere, lo cual indica que existen comunas mucho más caras que otras, destacando Las Condes, donde la caída del valor del reembolso llega hasta 1 UF, lo que en otras comunas ya podría ser clasificado como un outlier. En relación con las otras comunas, cabe destacar que la comuna de Santiago es la que tiene una media menor, lo que indica que es una comuna más barata para la compañía. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis (tabla 28), que, bajo una significancia del 0.05, indica que el reembolso aplicado sí varía dependiendo de la comuna.

4.2.3.7. Género

Corresponde a la variable que indica el género del titular del seguro, es una variable de interés, puesto que hombres y mujeres tienen un uso distinto del seguro y distintas responsabilidades que pueden afectar su utilización. La cartera se compone de la siguiente manera (figura 23):

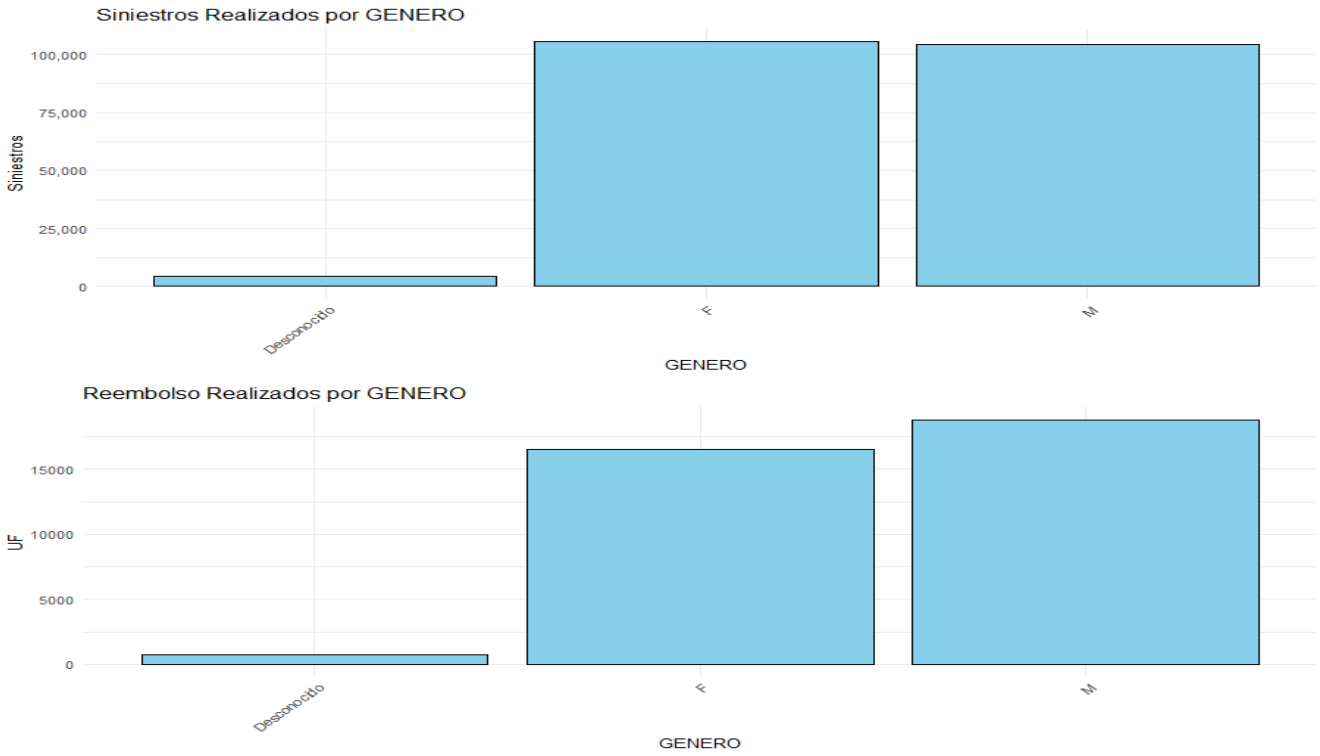

Figura 23: Distribución por Categoría de Rubro Económico

Se observa una proporción similar de siniestros en los grupos de hombres y mujeres, siendo la cantidad de siniestros ligeramente mayor en mujeres. Sin embargo, lo contrario ocurre con el reembolso, donde es ligeramente mayor en hombres. Además, se añade una categoría cuyo género se tiene como no identificado, cuyo impacto es muy bajo.

Número de Siniestros: Se estudiará el comportamiento de las variables de estudio por categorías de Genero(figura 24 tabla29):

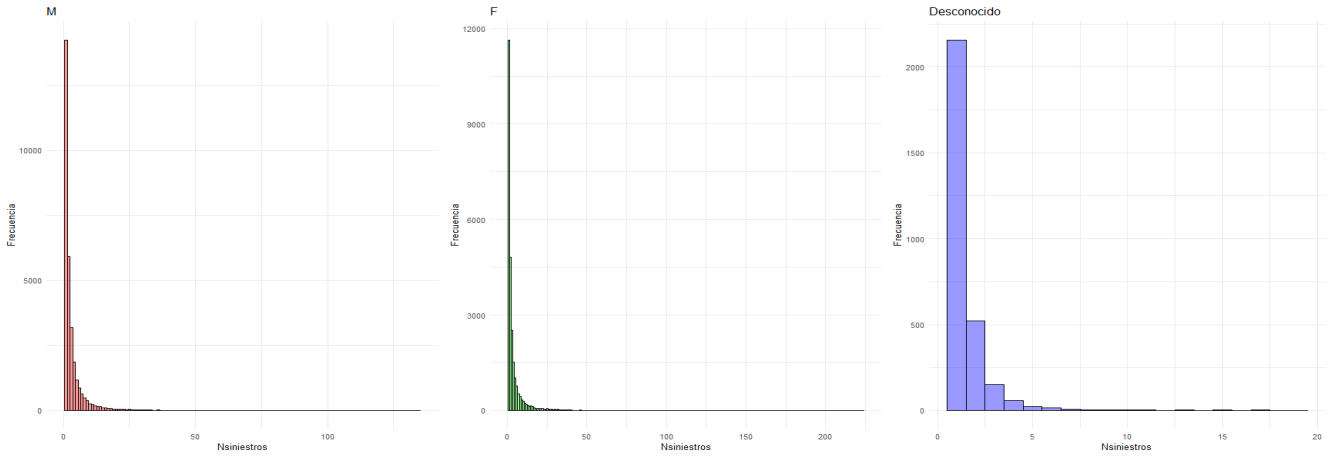

Figura 24: Distribución del número de Siniestros por Categoría de Género

Tabla 29: Resumen Descriptivo de Siniestros por Categoría de Genero

Genero	n	minimo	Q1	mediana	media	varianza	Q3	maximo
Desconocido	2954	1	1	1	1.49	1.56	2	19
F	25900	1	1	2	4.08	63.89	4	224
M	30715	1	1	2	3.4	33.98	3	135

Tabla 30: Resultado de Kruskal-Wallis para Siniestros por Categoría de Genero

Kruskal-Wallis	χ^2=40.87, df=1, p=1.6264e-10

Se observa que los titulares femeninos (tabla 30) poseen una mayor cantidad de outliers, además de tener una media y varianza mayor que los titulares masculinos, lo que puede indicar que se comportan de manera diferente. Para verificar lo anterior, se aplica el test de Kruskal-Wallis, que, bajo la hipótesis nula planteada en 30, se obtiene que, bajo una significancia del 0.05, el número de siniestros depende del género del titular.

Reembolso Aplicado: En relación con el comportamiento del reembolso aplicado (figura 25 y tabla 31), se observa lo siguiente:

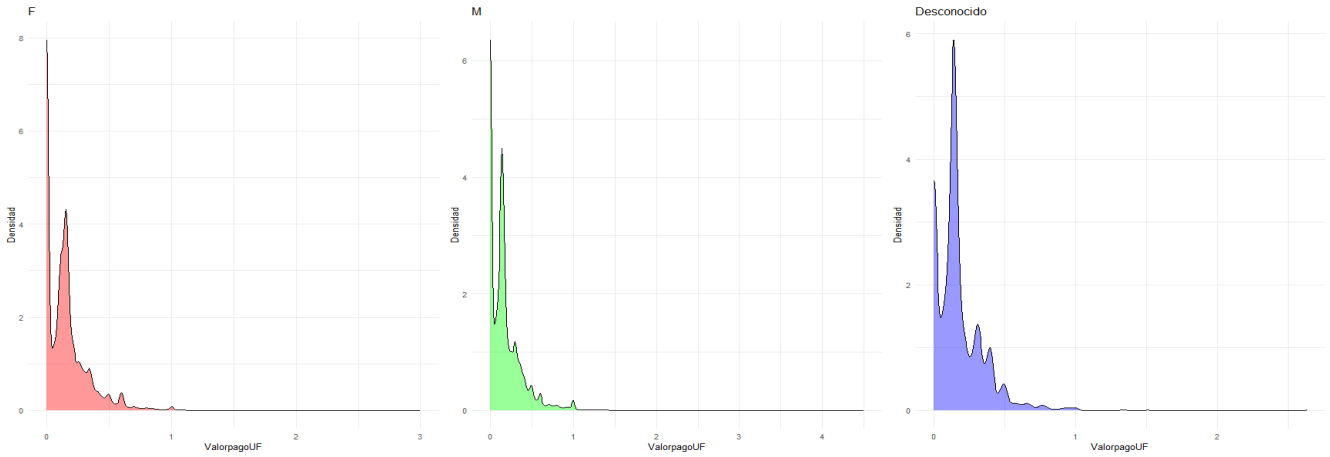

Figura 25: Distribución del Reembolso por Categoría de Género

Tabla 31: Resumen Descriptivo de Reembolso por Categoría de Género

Género	n	Mínimo	Q1	Mediana	Media	Varianza	Q3	Máximo
Desconocido	4394	0	0.08	0.14	0.18	0.03	0.24	2.63
F	105571	0	0	0.13	0.16	0.03	0.21	3
M	104519	0	0.01	0.14	0.18	0.04	0.25	4.5

Tabla 32: Resultado de Kruskal-Wallis para Reembolso por Categoría de Género

Kruskal-Wallis	χ^2=470.32, df=1, p=< 2.22e-16

En este caso, se observa que las distribuciones difieren visualmente (figura 31). Los titulares masculinos están asociados con siniestros cuyos valores son más cercanos a cero, en cambio, las mujeres tienen una mayor cantidad de siniestros cuyo reembolso es cero para la compañía. Para estudiar la diferencia entre los grupos, se aplica el test de Kruskal-Wallis(tabla 32), que, bajo una significancia del 0.05, indica que el reembolso sí varía dependiendo del género del titular.

4.2.3.8. Por Periodo de Liquidación

Corresponde a la combinación de las variables mes y año, se añade con el objetivo de estudiar las diferencias de cómo se comportan estas variables a lo largo del tiempo, el cual se comporta de la siguiente forma (figura 26):

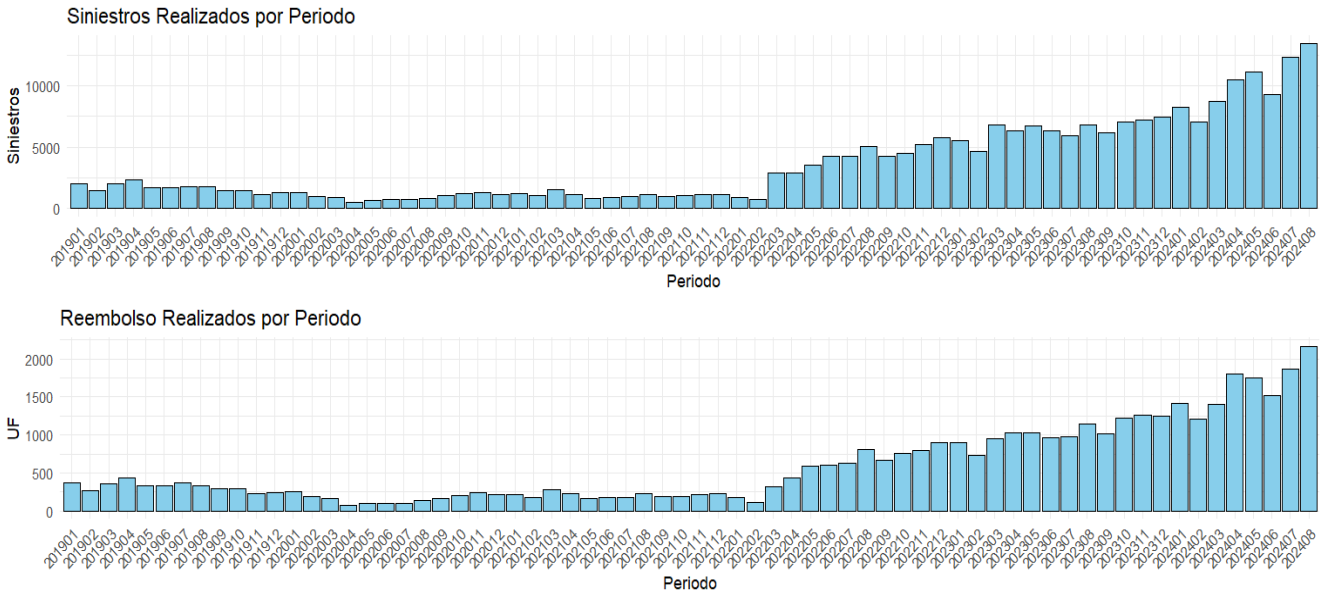

Figura 26: Distribución de Siniestros Iniciales por Categoría de Periodo de Liquidación

Donde se observa un desbalance, ya que a lo largo del tiempo se suman cada vez más pólizas a la compañía, lo que incrementa la cantidad de siniestros a reembolsar. Si se quiere observar la diferencia entre los meses, podemos ver lo siguiente:

4.2.4. Distribución de Variables Continuas

4.2.4.1. Valor de la Prestación

Corresponde al valor original del servicio prestado, definido tanto por el tipo de servicio como también por el prestador de salud en el que dio la atención. Esta se comporta de la siguiente forma (figura 27 tabla 33):

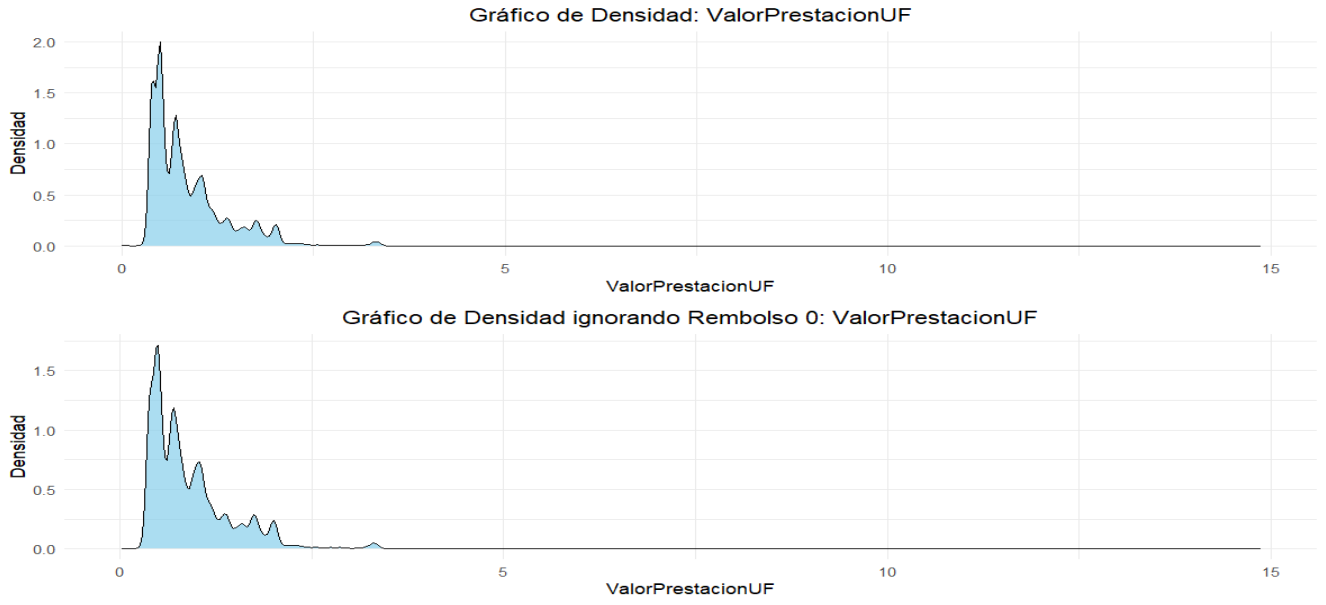

Figura 27: Distribución del Valor de la Prestación

Tabla 33: Estadísticas descriptivas de Valor Prestación

Vector	Media	Mediana	Desv. Est.	Mínimo	Máximo	Cuartil 25 %	Cuartil 50 %	Cuartil 75 %	Coef. Var.	Curtosis
Valor Prestación	0.8685063	0.6930	0.5872750	0.0000	14.8671	0.4845	0.6930	1.0592	67.61897	74.72449
Mayor que 0	0.9183465	0.7397	0.6188048	0.0302	14.8671	0.4880	0.7397	1.1165	67.38249	71.21529

Para medir correctamente si existe diferencia dentro de la variable cuando el reembolso aplicado es mayor a cero, se creo un subconjunto con esta condición, encontrándose que no se ven diferencias a simple vista en ambos casos y comparando su comportamiento bajo la tabla 33, se observa que las diferencias son mínimas, por lo cual se el uso de esta variable no ayuda a distinguir entre los casos.

4.2.4.2. Valor de Bonificación Primera Capa

Corresponde al valor que fue reembolsado por la Isapre o Fonasa del titular, el cual se comporta de la siguiente manera (figura 28 tabla 34):

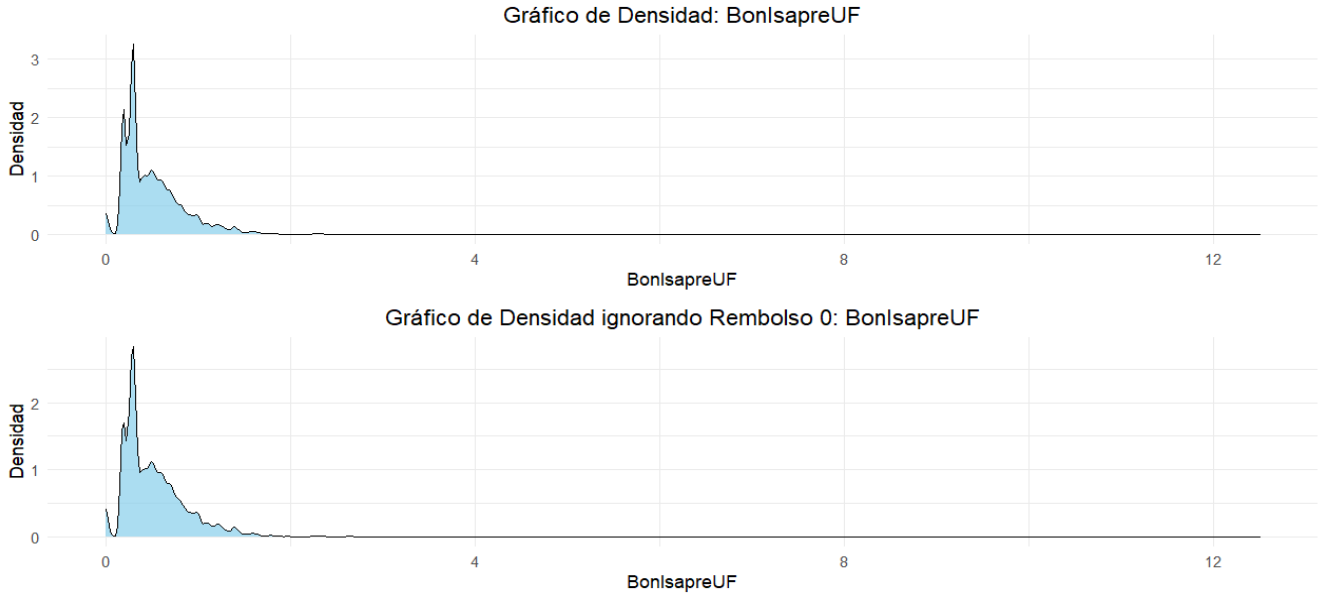

Figura 28: Distribución del Valor de la Bonificación de la Primera Capa

Tabla 34: Estadísticas descriptivas de Valor Bonificación de la Primera Capa

Variable	Media	Mediana	Desv. Est.	Mínimo	Máximo	Cuartil 25 %	Cuartil 50 %	Cuartil 75 %	Coef. Var.	Curtosis
Bonificación	0.5052517	0.3996	0.3995453	0	12.5126	0.2779	0.3996	0.6551	79.07847	120.2196
Mayor que 0	0.5246865	0.4266	0.4154094	0	12.5126	0.2841	0.4266	0.6872	79.17287	119.5037

En este caso, podemos observar que la distribución no muestra diferencias significativas al comparar con la distribución cuando el reembolso es positivo. Lo cual lleva a la conclusión de que, el reembolso puede ocurrir independientemente de la bonificación de la Isapre. Al observar las estadísticas descriptivas(tabla 34), se confirma lo que se observa gráficamente.

4.2.4.3. Valor del Copago

Corresponde a la diferencia entre el valor de la prestación y la bonificaron de la primera capa y algún seguro asociado que tenga el titular, sobre este valor se calcula el reembolso final y se comporta de la siguiente manera (figura 29 tabla 35):

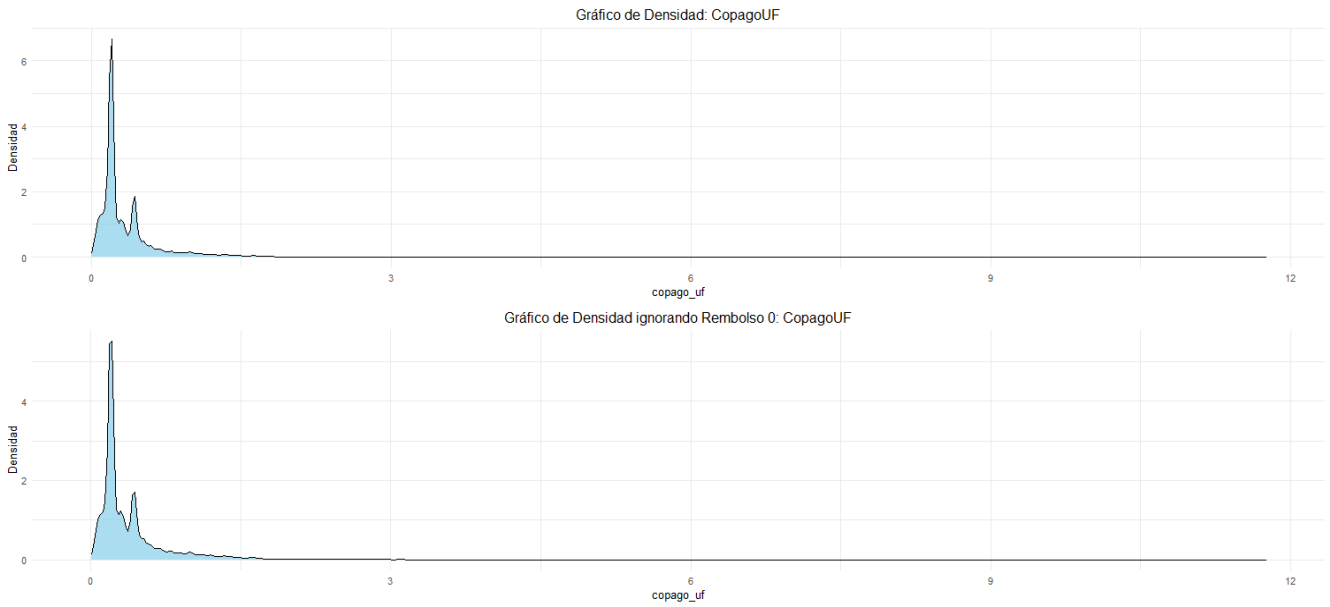

Figura 29: Distribución del Valor del Copago Aplicado

Tabla 35: Estadísticas descriptivas del valor del Copago

Variable	Media	Mediana	Desv. Est.	Mínimo	Máximo	Cuartil 25 %	Cuartil 50 %	Cuartil 75 %	Coef. Var.	Curtosis
Valor Copago	0.3589	0.2127	0.3570	0.0000	11.7586	0.1917	0.2127	0.4280	99.4717	33.1961
Mayor que 0	0.3889	0.2296	0.3831	0.0027	11.7586	0.1929	0.2296	0.4346	98.5158	28.8174

Se observa en 29 que tiene una distribución similar a una normal, con altos outliers. Cabe destacar que esta distribución no cambia cuando el reembolso es cero, por lo tanto, al estudiar la separación entre el cero y el resto de los valores, no ofrece mucho valor.

4.2.4.4. Valor del Deducible Aplicado

Según Fundación MAPFRE (2025) el deducible es la cantidad o porcentaje establecido en una póliza cuyo importe ha de superarse para que se pague una reclamación. En este caso corresponde al deducible que le fue aplicado al titular para poder recibir un reembolso, el cual se comporta de la siguiente manera(figura tabla):

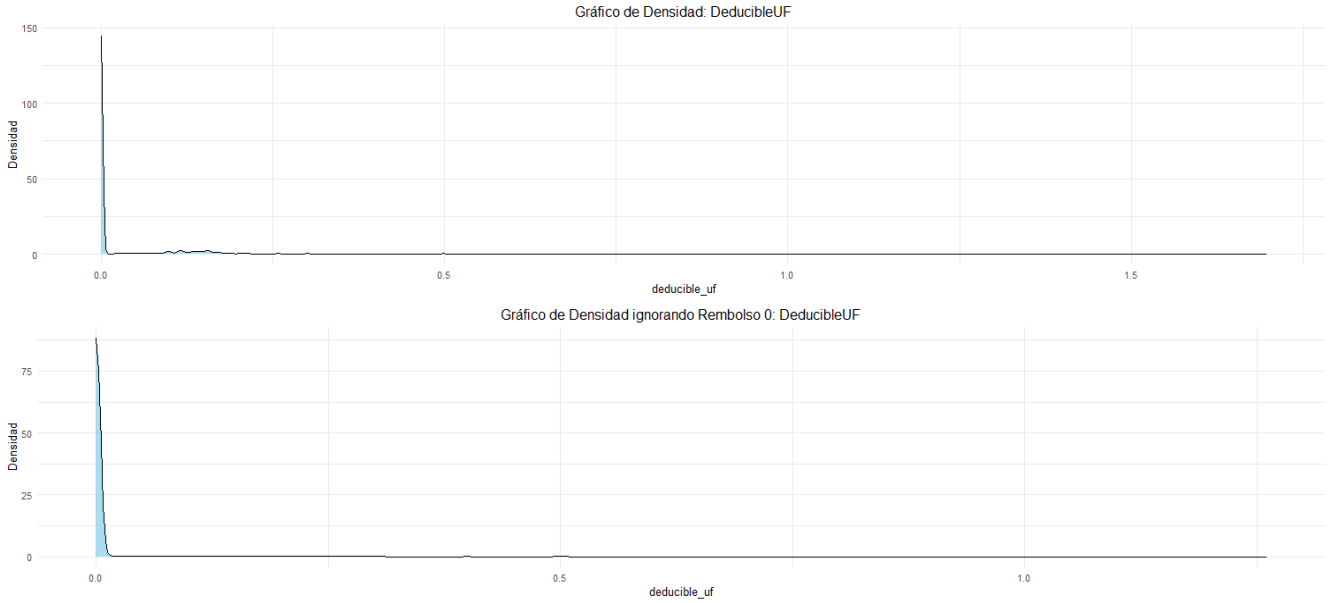

Figura 30: Distribución del Valor del Deducible Aplicado

Tabla 36: Estadísticas descriptivas del Valor del Deducible Aplicado

Variable	Media	Mediana	Desv. Est.	Mínimo	Máximo	Cuartil 25 %	Cuartil 50 %	Cuartil 75 %	Coef. Var.	Curtosis
Valor del Deducible	0.045403563	0	0.09834499	0	2.6443	0	0	0.0467	216.6019	29.12937
Mayor a 0	0.008174212	0	0.04994472	0	1.5000	0	0	0.0000	611.0035	100.72610

Se puede observar que la distribución está muy acotada en torno al cero. Esto ocurre porque el deducible se impone en la póliza para desincentivar el uso del seguro, puesto que impone un pago mínimo para recibir reembolso, si comparamos el comportamiento cuando existe reembolso, se observa que cuando no hay reembolso, hay un aumento de la media y de los cuantiles, también considerando que los outliers se concentran en este caso, por lo que hay sospechas que esta mucho mas relacionado cuando no hay reembolso.

4.2.5. Correlación

En esta sección se estudiará la correlación entre las variables numéricas usadas en el estudio, además de las técnicas aplicadas en las variables en caso de presentar problemas asociados con la multicolinealidad de la información.

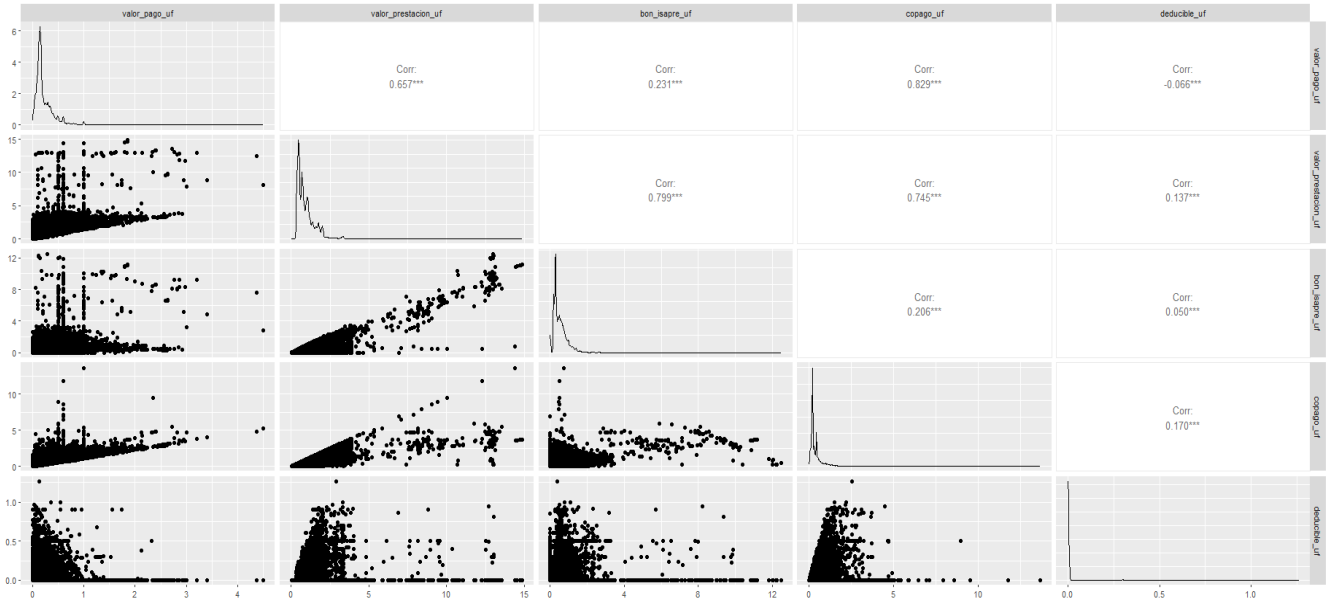

Figura 31: Correlacion Variables Continuas

De la cual, podemos observar (figura 31) que todas las variables introducidas tienen una alta correlación con el valor del reembolso (ValorPagoUF), excepto por deducible, cuya correlación es baja, considerando que por el momento solo se consideran los valores mayores que 0. Además, cabe mencionar que las correlaciones encontradas tienen carácter lineal, pero se observa que muchas de ellas tienen observaciones que no siguen la tendencia esperada. Donde el caso más interesante se encuentra con deducible, puesto que se forma una cota en los valores que puede tomar, pero entre más cercano sea el valor a cero, más se infla el valor del deducible, lo que ya muestra una fuerte tendencia a diferenciar un reembolso positivo.

Por lo tanto, el conjunto de variables continuas a utilizar quedará con las restantes para estudiar los casos de reembolsos mayores que cero, y utilizar el deducible para poder distinguir entre ambos casos. Pero, entre las variables predictoras, se puede observar que existe una alta correlación entre ellas, lo que puede implicar problemas de multicolinealidad y afectar las interpretaciones de los parámetros y el desempeño del modelo. Por lo que se estudiará la viabilidad de realizar una reducción de dimensionalidad utilizando el método de componentes principales.

4.2.6. Componentes Principales

Como se observó anteriormente, las variables predictoras presentan una alta correlación entre sí, por lo que es necesario estudiar si se puede desarrollar un conjunto de menor dimensión de variables que entregue la misma información y cumpla con el criterio de no multicolinealidad. Para ello, se debe estudiar si efectivamente el conjunto de variables presenta las características para ser viable.

4.2.6.1. Viabilidad del Estudio

Se realizaro un test de esfericidad de Bartlett, el cual puede ser revisado en A.3, cuyo objetivo es verificar si el determinante de la matriz de correlaciones del conjunto de variables presenta correlaciones significativas para realizar un estudio, obteniendo los siguientes resultados (tabla 37).

Tabla 37: Resultados del Test de Esfericidad de Bartlett

Estadístico	**Valor**
Chi-cuadrado (χ^2)	780.02
p-valor (p)	< 0.0001
Grados de libertad (df)	6

Podemos concluir que el conjunto sí presenta suficiente correlación para justificar el uso de técnicas como el PCA. Para complementar el anterior resultado, se calculó el índice KMO (planteado en A.4), con el objetivo de estudiar si la correlación que presenta nuestro conjunto de datos es suficiente para justificar el uso de componentes principales, cuyo resultado se observa en la siguiente tabla:

Tabla 38: Resultados del Índice Kaiser-Meyer-Olkin (KMO)

Variable	**MSA (Adecuación)**
ValorPrestacion	0.38
BonIsapre	0.26
Deducible	0.99
Copago	0.25
KMO General	0.31

De los resultados de la tabla anterior (tabla 38) , el índice KMO es bajo e insuficiente para el uso de técnicas como el PCA. Esto puede suceder debido a que las correlaciones parciales son mucho más predominantes que las correlaciones simples entre ellas, lo que provoca que los componentes obtenidos no sean los idóneos. Al observar cuáles son las variables más problemáticas para el estudio, se observa que el deducible, bajo este criterio, está muy correlacionado con el resto de variables, por lo que su uso debe ser medido al estar en presencia de otras variables.

4.2.7. Resultados de la Modelación

Inicialmente, antes de aplicar los modelos, se realiza una partición de la información. Se crean dos sets para cada base de datos: un set de entrenamiento, el cual comprende toda la información de la base de datos hasta los últimos cuatro meses del último año registrado, y el set de testeo, el cual tiene la información restante, con el objetivo de finalmente poder estudiar el poder predictivo del modelo con datos de meses desconocidos.

Tabla 39: Tabla 4x4 de Entrenamiento y Testeo

Base/Set	**Entrenamiento**	**Testeo**
Reembolso	196.687	49.171
NSiniestros	41.288	10.308

4.2.8. Modelación del Reembolso Aplicado

Como se mencionó anteriormente, el objetivo de esta etapa es modelar los costos directos de la compañía y obtener los patrones e índices más relevantes. Sin embargo, según lo observado en el análisis descriptivo, existe una gran cantidad de ceros que no contribuyen al valor real de los costos.

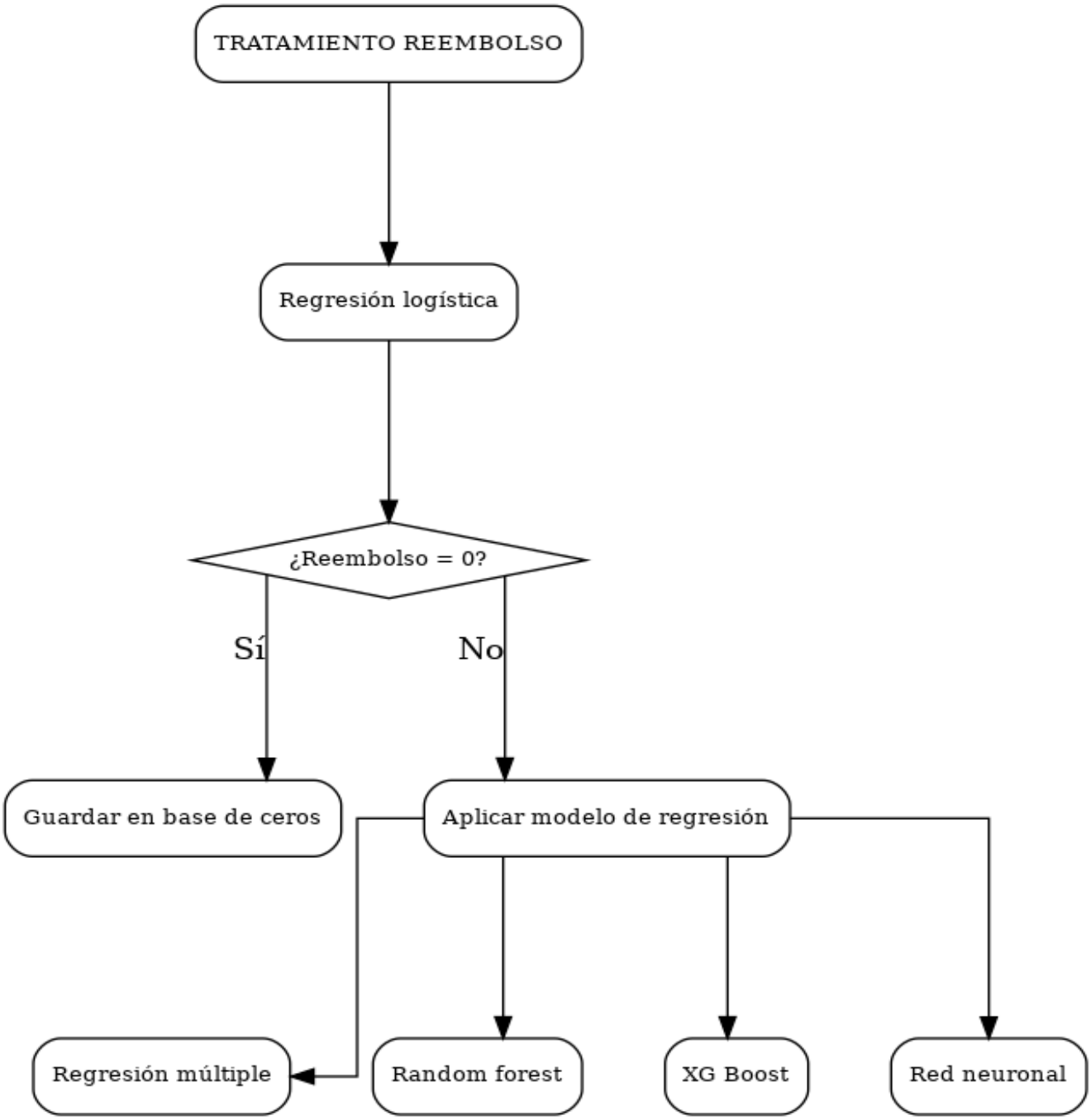

Figura 32: Desarrollo de la Modelación del Reembolso Aplicado

La figura anterior 32 describe el esquema de trabajo propuesto para tratar con el problema de una distribución con ceros inflados, la cual se divide en dos etapas:

- En la primera etapa es separar la inflación de ceros del resto de las observaciones. Para ello, en primera instancia, se creará una variable ficticia de tipo binaria llamada Y_class y se usará la base para entrenar modelos clasificatorios y realizar una partición correcta.
- En la segunda etapa la segunda el objetivo es entrenar modelos de regresión para poder generar futuras predicciones y encontrar relaciones entre los los datos.

4.2.8.1. Etapa 1

En esta etapa se planteó una variable llamada Y_class, la cual divide la base en dos, donde toma valor 0 cuando el reembolso aplicado es menor a 0.001 UF, esto para agrupar todos los reembolsos que tengan valor real menor a 500 pesos chilenos. La distribución se comporta de la siguiente manera (tabla 40):

Tabla 40: Frecuencia de Clases en `y_class`

Clase	**Frecuencia**
0	55,220
1	159,254
Total	214,474

Para todos los modelos se utilizó inicialmente el mismo conjunto de variables:

Tabla 41: Lista de variables categóricas y continuas para la respuesta binaria.

Categóricas	**Continuas**
Año	Deducible
Mes	Bonificación
Comunas	Valor prestacion UF
Región	Copago UF
Relación	
Servicio	
Primera capa(IAgrupado)	
Prestadores	
Género	

4.2.8.1.1. Regresión Logística

Se utiliza como modelo base una regresión logística, la cual es de interés por la capacidad de brindar interpretaciones que, a diferencia de los modelos de machine learning, no permiten una interpretación directa. Se aplicó el modelo con todas las variables y se redujo mediante el método stepwise hasta obtener el siguiente modelo reducido, en el cual solo se muestran las variables que tienen mayor peso a la hora de discriminar entre las dos clases. El modelo completo puede observarse en B.4.

Tabla 42: Resultados del Modelo Logístico con Odds Ratios (Mayores Impactos)

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**Odds Ratio**
DeducibleUF	-29.76592	0.14923	-199.457	$< 2e-16$	6.83×10^{-14}
CopagoUF	8.76669	0.13795	63.563	$< 2e-16$	6431.61
Rel_Hijo.a	-0.34921	0.03188	-10.958	$< 2e-16$	0.7054
Género_F	-0.23686	0.02379	-9.954	$< 2e-16$	0.7889
Género_M	-0.25632	0.02392	-10.719	$< 2e-16$	0.7738
Servicio_Dermatología	-0.34169	0.05963	-5.729	$1.00e-08$	0.7106
Servicio_Urología	0.16015	0.07655	2.092	0.0364	1.1737

Tabla 43: Métricas de ajuste del modelo logístico

Métrica	**Valor**
Null deviance	198438
Residual deviance	93241
AIC	93343

Se observa en la tabla 42 que la variable deducible es la que tiene mayor peso a la hora de discriminar, lo cual reafirma lo observado en el análisis descriptivo. Por cada unidad de aumento en el deducible, disminuye la probabilidad de que el reembolso sea mayor a cero en 10^{-14}. Además, cabe destacar que la variable copago tiene un gran impacto, ya que por cada unidad de aumento, aumenta la probabilidad de que el reembolso sea mayor a cero en 6.431. Cabe destacar también que el género, en ambos casos, disminuye la probabilidad en comparación a que este sea desconocido, y que las consultas dermatológicas disminuyen la probabilidad de que el reembolso sea aplicado.

Rendimiento del Modelo: Se realiza clasificación del set de testeo utilizando el modelo expuesto bajo un punto de corte de 0.5 y se compara con los valores reales, los cuales se reflejan en las tablas 44 y 45:

Tabla 44: Tabla de Confusión del Modelo Logístico

Predicción	**Real = 0**	**Real = 1**	**Total por Predicción**
0	8192	560	8752
1	1910	31153	33063
Total por Real	10102	31713	41815

Tabla 45: Métricas de Desempeño del Modelo Logístico

Métrica	**Valor**
Accuracy	0.9360756
Precision	0.9518578

El rendimiento del modelo es adecuado, ya que se observa en la tabla 44 una baja tasa de falsos positivos y falsos negativos, lo que es una buena señal considerando la escala de los datos presentados. Esto se respalda con las métricas observadas en 45, donde los valores están muy cerca de uno, lo que indica que la clasificación es casi perfecta. Sin embargo, se observa una tendencia a clasificar valores cero como si fueran uno, lo que sugiere que aún hay margen de mejora en el modelo.

Curva ROC: Para medir el rendimiento del modelo en términos de sensibilidad y especificidad, y determinar el punto de corte óptimo del modelo, se procede a estudiar su curva ROC (figura 33), la cual está dada por:

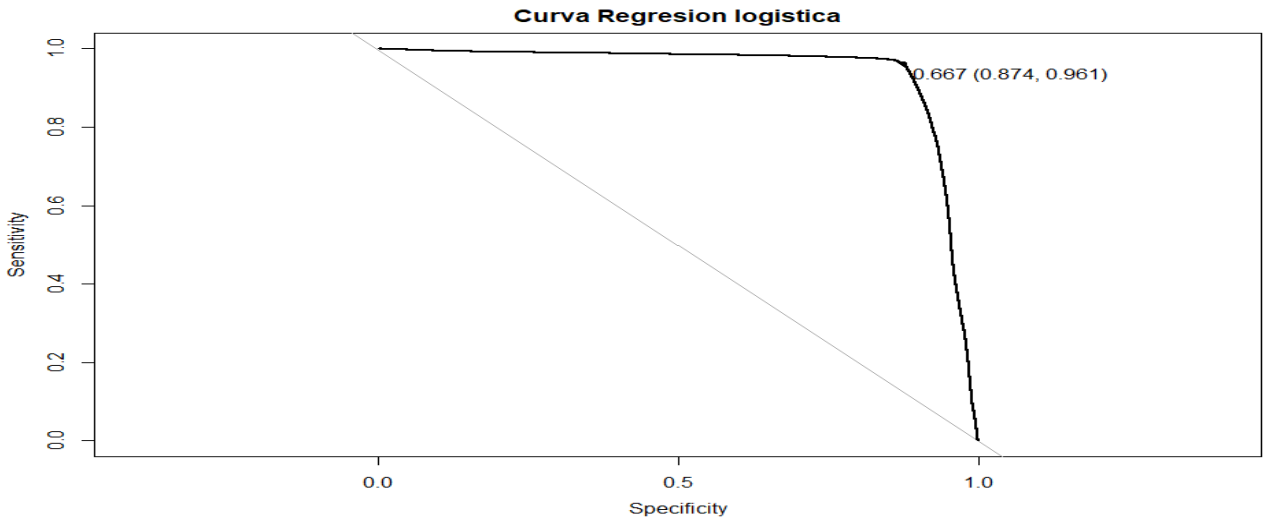

Figura 33: Curva ROC del Modelo Logístico

La curva nos indica que el punto de corte óptimo, que maximiza tanto la sensibilidad como la especificidad, es de 0.667, con un área bajo la curva de 0.92686. Con esta información, se vuelve a estudiar el rendimiento del modelo para evaluar las mejoras que implica este ajuste en las tablas 46 y 47:

Tabla 46: Tabla de Confusión del Modelo Logístico con el Punto de Corte Óptimo

Predicción	**Real = 0**	**Real = 1**	**Total por Predicción**
0	8555	1126	9681
1	1547	30587	32134
Total por Real	10102	31713	41815

Tabla 47: Métricas de Desempeño del Modelo Logístico con el Punto de Corte Óptimo

Métrica	**Valor**
Accuracy	0.9409303
Precision	0.9422315

Se observa que el punto de corte óptimo mejora el problema encontrado anteriormente(tabla 44), donde la tasa de falsos positivos era más alta de lo esperado. Sin embargo, al ajustar el punto de corte (tabla 46), se aumenta la tasa de falsos negativos, lo cual es esperado debido a la escala de los datos, ya que aún existen observaciones muy cercanas al punto de corte propuesto. Finalmente, los resultados obtenidos generan un debate sobre si utilizar o no el punto de corte óptimo.

4.2.8.1.2. Modelos de Redes para Clasificación

Dado que se trata de un problema de clasificación, se tomó la decisión de probar el ajuste del modelo bajo redes neuronales. Aunque se pierde la capacidad de interpretabilidad del modelo anterior, se espera un modelo mucho más robusto. La arquitectura completa se puede encontrar en B.1, el cual fue entrenado bajo los siguientes hiperparametros(tabla 48):

Tabla 48: Hiperparámetro Redes Neuronales Regresión

Hiperparámetro	Valor
Dropout	0.4
Batch size	128
Learning rate	10^{-4}

Evaluación del Modelo Inicialmente, se presentan las métricas asociadas al modelo junto con su tabla de confusión (tablas 49 y 50):

Tabla 49: Tabla de Confusión de Redes Neuronales

Predicción	Real = 0	Real = 1	Total por Predicción
0	6777	340	7117
1	3325	31373	34798
Total por Real	10102	31713	41815

Tabla 50: Métricas de Desempeño del Modelo de Redes Neuronales

Métrica	Valor
Accuracy	0.912352
Precision	0.9041732

Donde el desempeño es menor al obtenido en 4.2.8.1.1., ya que se observan más observaciones mal clasificadas como mayores que cero. A diferencia del modelo anterior, es muy difícil que este modelo clasifique mal las observaciones que son efectivamente mayores que cero.

Curva ROC: Se presenta la curva ROC para medir el desempeño en términos de sensibilidad y especificidad(figura 34):

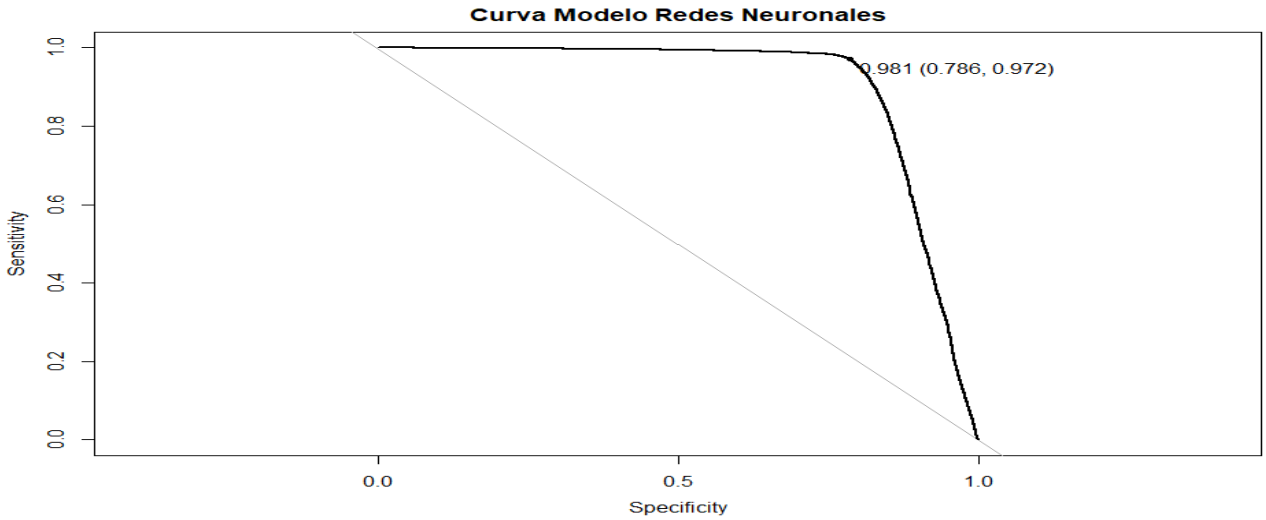

Figura 34: Curva ROC del Modelo de Redes Neuronales

En la curva ROC podemos observar que el punto de corte óptimo para estos datos es de 0.981, con una sensibilidad de 0.786 y una especificidad de 0.972. La especificidad es ligeramente mayor que la obtenida

en el modelo logístico, pero con un mayor costo computacional. A continuación, se muestra la tabla de confusión obtenida con este modelo(tablas 51 y 52):

Tabla 51: Tabla de Confusión del Modelo de Redes Neuronales, Punto de Corte Óptimo

Predicción	**Real = 0**	**Real = 1**	**Total por Predicción**
0	8289	2533	10822
1	1813	29180	30993
Total por Real	10102	31713	41815

Tabla 52: Métricas de Desempeño del Modelo de Redes Neuronales, Punto de Corte Óptimo

Métrica	**Valor**
Accuracy	0.896066
Precision	0.9415029

Del cual, en términos de métricas, se obtuvo una mejora en la precisión, es decir, aumentó la tasa de verdaderos positivos, a cambio de una pérdida de precisión, lo que indica que el modelo tiene un menor poder predictivo en general. Por lo tanto, el modelo logístico sigue siendo una mejor opción.

Importancia de Variables: Finalmente, al valorar las variables que más aportan al modelo, se encontró lo siguiente (figura 35):

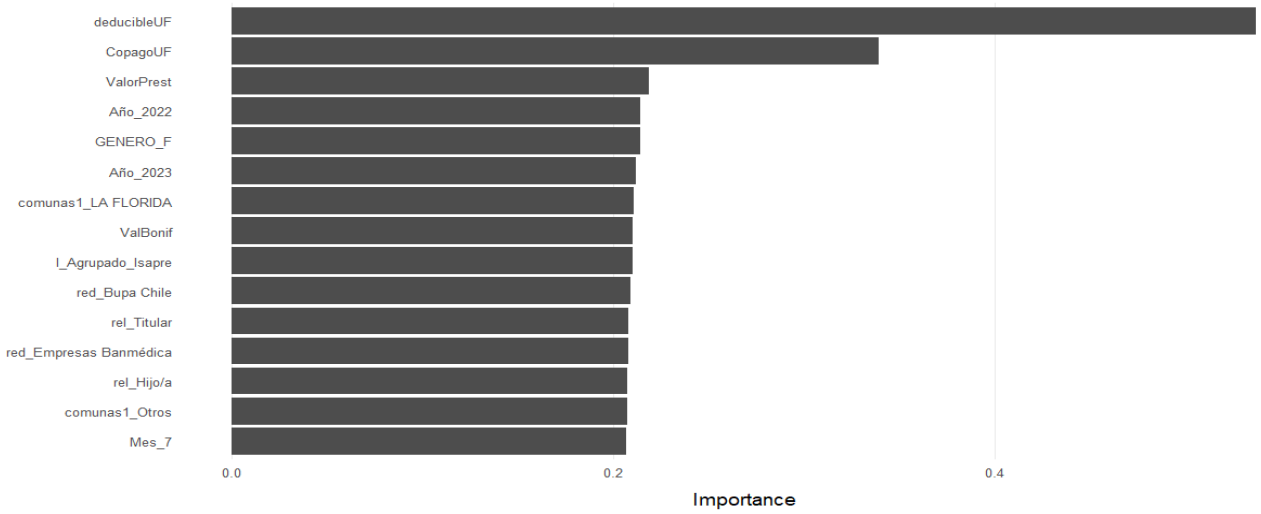

Figura 35: Importancia de Variables en el Modelo de Redes Neuronales y Logístico

Se puede destacar que "deduciblez çopago"son las variables que tienen mayor peso, repitiéndose lo obtenido en el modelo logístico. También se observa que ser titular tiene un mayor peso en la predicción, especialmente cuando está asociado a una póliza de Isapre.

4.2.8.2. Etapa 2

En esta etapa, una vez diferenciado cuándo la compañía realiza un reembolso mayor a 0.0001, el siguiente paso es determinar un modelo capaz de predecir el reembolso esperado, basándose en las relaciones existentes

con las variables predictoras estudiadas en el análisis descriptivo. Este modelo se llevará a cabo utilizando los enfoques planteados en 4.2.7. Las variables seleccionadas inicialmente serán las siguientes:

Tabla 53: Lista de variables categóricas y continuas para el modelo de Reembolso

Categóricas	**Continuas**
Año	Copagouf
Mes	Bonificaciónuf
Comunas	ValorPrestación
Reg	Deducibleuf
Relación	
Género	
Primera capa (IAgrupado)	
Prestadores	
Servicio	

4.2.8.2.1. Modelo de Regresión Múltiple

Como se observo en el análisis descriptivo, el reembolso aplicado, tiene una distribución similar a una normal, con una cola superior muy pesada, por lo que es de interés desarrollar un modelo que sirva de base para entender la mejora que se encontrara mas adelante con los modelos de machine learning. Se desarrolla el modelo con todas las variables y se muestra el modelo reducido en 2 y se como forma de resumen se muestran las variables mas importantes para el modelo(tabla 54):

Tabla 54: Variables más importantes RLM

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
deducibleuf	-0.8584	0.0052	-164.846	$< 2 \times 10^{-16}$
Copagouf	0.5313	0.0044	119.619	$< 2 \times 10^{-16}$
Red_Condes	0.0652	0.0019	34.925	$< 2 \times 10^{-16}$
Red_Grupo.Alemana	0.0700	0.0022	31.145	$< 2 \times 10^{-16}$
comunas_LAS.CONDES	0.0345	0.0012	28.143	$< 2 \times 10^{-16}$

De la tabla anterior 54, podemos observar que las variables más importantes son el deducible, puesto que por cada unidad de aumento, la media del reembolso disminuye en 0.8584, y el copago, donde por cada unidad de aumento de esta variable, la media del reembolso aumenta en 0.5313, manteniendo constantes el resto de las variables. Sin embargo, cabe destacar que la escala del deducible implica que su impacto es muy bajo en comparación con el copago. También se destaca que las variables que implican un aumento de la media de reembolsos son estar en prestadores de Las Condes y del Grupo Alemana, así como también atenderse en Las Condes, lo cual resulta más caro que en otras comunas.

Validación de Supuestos: A continuación (tabla 55) se aplican los test necesarios para estudiar los supuestos del modelo y poder validar las inferencias y conclusiones que se tomaron:

En este caso, podemos observar que los supuestos estudiados no se cumplen, lo que implica que los resultados de los modelos no son completamente fiables para realizar inferencias. Sin embargo, con el objetivo de poder comparar el rendimiento de este modelo con el de otros, se llevarán a cabo pruebas de su capacidad predictiva.

Tabla 55: Resultados de los supuestos RLM

Supuesto	**Prueba**	**Resultados**
Independencia de los errores	Durbin-Watson	D-W Statistic = 1.1965, p-value < 2.2e-16
Homocedasticidad	Breusch-Pagan	BP = 26699, df = 48, p-value < 2.2e-16
Normalidad de los errores	Anderson-Darling	A = 6466.7, p-value < 2.2e-16

Tabla 56: Métricas de Evaluación para RLM

Métrica	**Valor**
MAE (Mean Absolute Error)	0.052
MSE (Mean Squared Error)	0.0094
MAPE (Mean Absolute Percentage Error)	38.07 %
R^2 (Coeficiente de Determinación)	0.7185

Podemos observar que a pesar que el modelo no es el optimo, las métricas indican que el modelo esta prediciendo los valores, pero el error es suficientemente alto para que estos no sean fiables.

Reales vs Predichos: Para evaluar el poder predictivo del modelo, se estudia su desempeño con la base de testeo (figura 36):

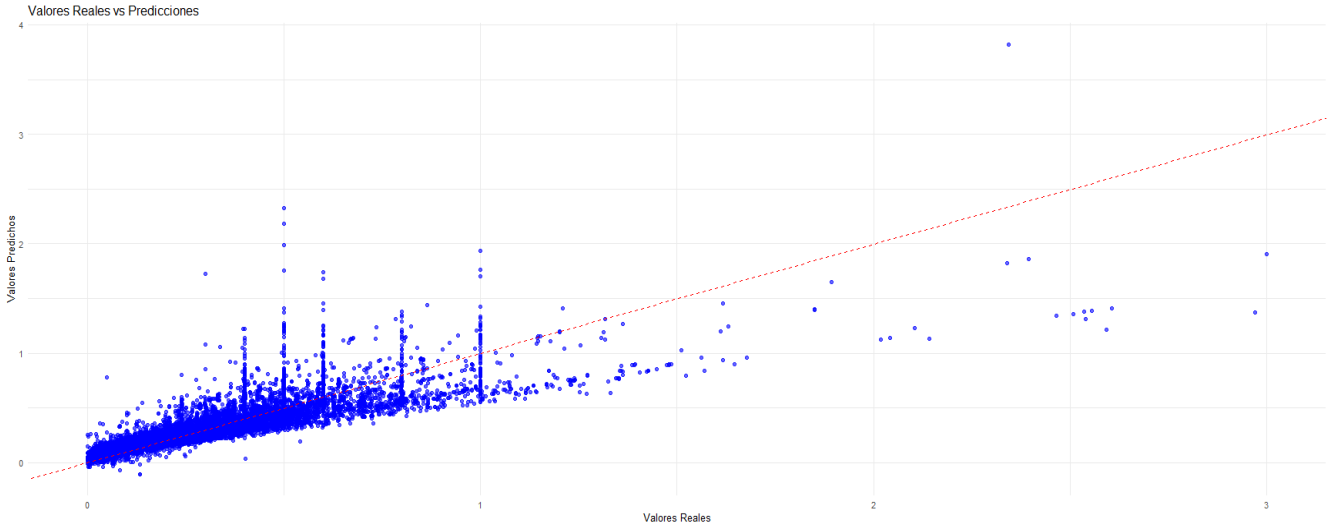

Figura 36: Gráficos Reales vs Predicciones Modelo de Regresión.

Observando el gráfico (figura 36), se observa que el modelo efectivamente está prediciendo los datos, pero hay observaciones que no logra predecir con precisión, ya que tiene una tendencia a subestimar las observaciones con valores altos y a sobreestimar observaciones con valores bajos.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera(figura 37):

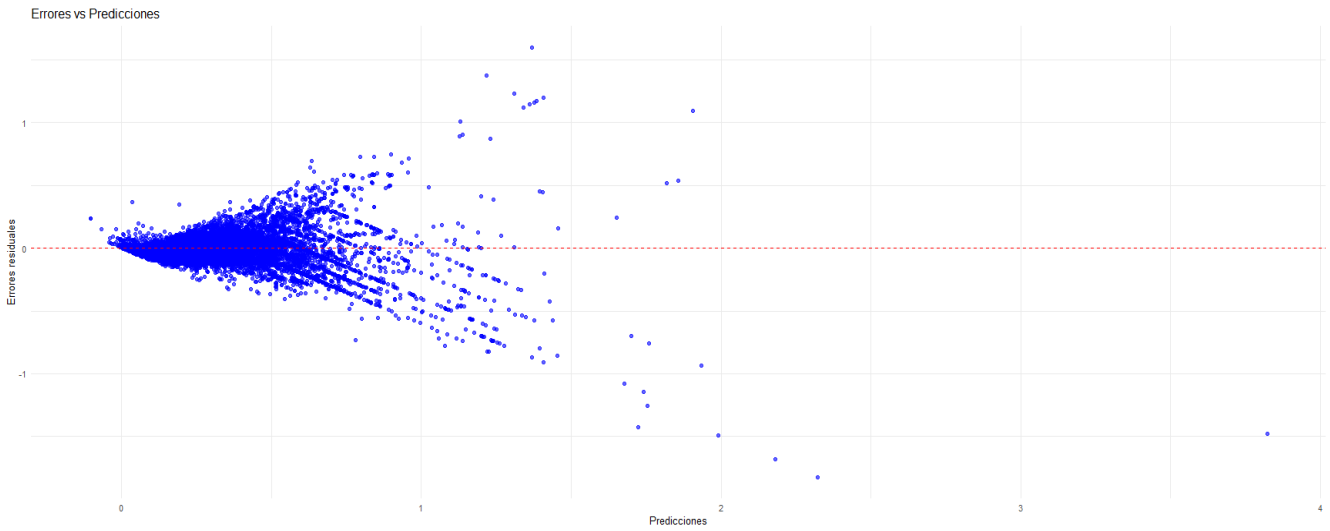

Figura 37: Gráficos Errores del RML

En este gráfico (figura 37) se observa que, acorde al test de Durbin-Watson, los errores son heterocedásticos, aumentando su valor absoluto a medida que aumenta el valor real. Además, el modelo tiende a subestimar, aumentando la magnitud de esta subestimación conforme aumenta el valor del reembolso.

4.2.8.2.2. Redes Neuronales

Dado que se trata de un problema de regresión, se tomó la decisión de probar el ajuste del modelo bajo redes neuronales. Donde se espera un mejor rendimiento y mayor ajuste comparado con el modelo anterior, la arquitectura usada puede ser encontrada en B.2 y los hiperparámetros usados ,para entrenar el modelo utilizado (tabla 57):

Tabla 57: Hiperparámetro Redes Neuronales Regresión

Hiperparámetro	**Valor**
Dropout	0.4
Batch size	128
Learning rate	10^{-4}

El modelo fue entrenado utilizando todas las variables disponibles y se muestran las métricas de rendimiento del modelo prediciendo los valores de la base de datos de testeo(tabla 58):

Tabla 58: Métricas de Desempeño del Modelo de Redes Neuronales Reembolso

Métrica	**Valor**
MAE (Mean Absolute Error)	0.0439
MSE (Mean Squared Error)	0.0069
MAPE (Mean Absolute Percentage Error)	26.80 %

Reales vs Predichos: Para evaluar el poder predictivo del modelo, se estudia su desempeño con la base de testeo (figura 38):

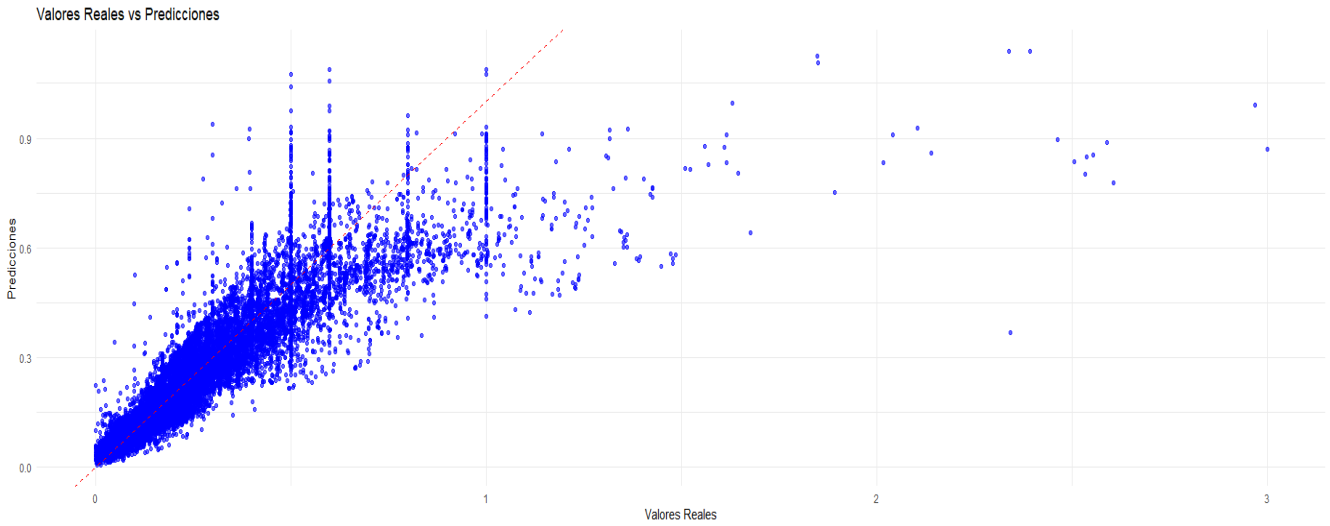

Figura 38: Gráficos Reales vs Predicciones Red Neuronal Reembolso

Donde podemos observar que el modelo, en los rangos bajos, es capaz de ajustar bien los datos, pero no es capaz de predecir los valores más altos, teniendo problemas mas marcados que el modelo RLM(4.2.8.2.1.), mostrando una alta tendencia a subestimar las observaciones. Esta tendencia aumenta a medida que el valor del reembolso crece.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera(figura 39):

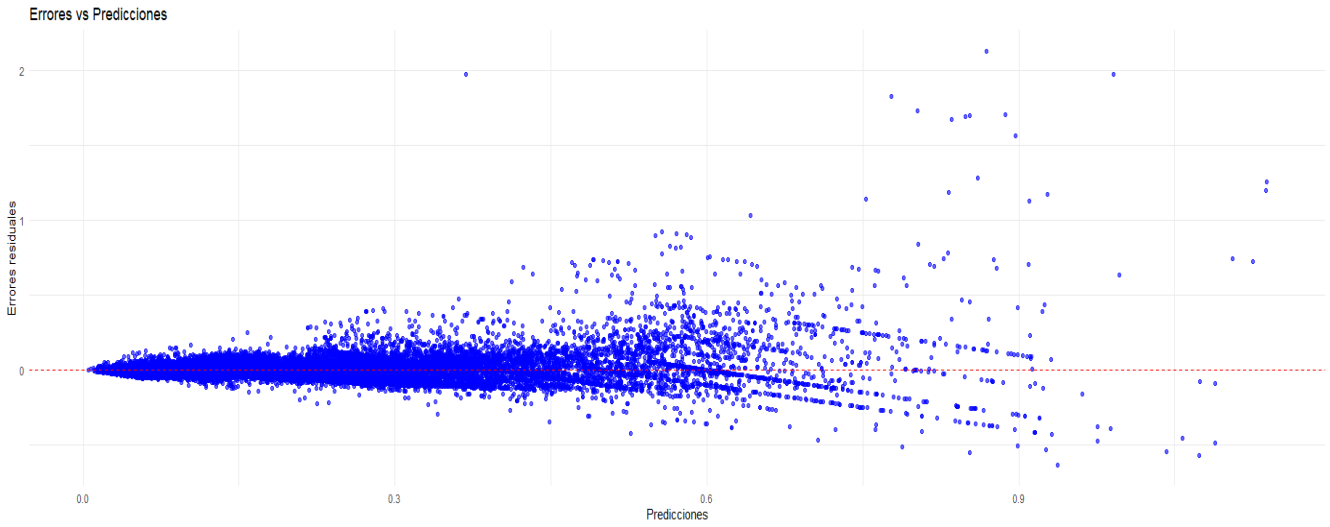

Figura 39: Gráficos de Errores Red Neuronal Reembolso

En general los errores son visualmente heterocedasticos y van aumentando en proporción al aumento de las predicciones, cabe destacar que la escala de los errores es baja pero teniendo en cuenta que la variable respuesta esta en uf, pequeños cambios impactan fuerte de manera monetaria. También se puede observar que el modelo tiene tendencia a subestimar el valor real del reembolso lo cual es un problema a la hora de querer ocupar este modelo en un problema real.

Importancia de Variables: Finalmente, a la hora de valorar las variables que más aportan al modelo, se pudo encontrar lo siguiente(figura 40):

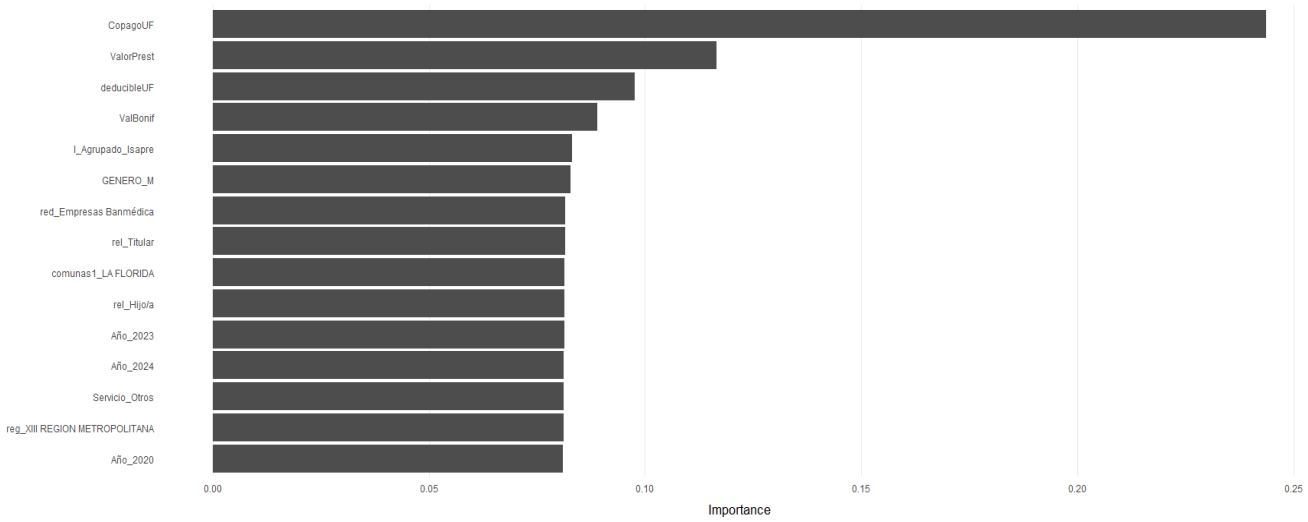

Figura 40: Gráficos de Importancia Redes Neuronales Reembolso

Se observa en 40 que la variable con mayor importancia es el copago. Si se elimina del modelo, se presenta un aumento del error o MSE del 0.2, lo cual es alto e implica que las predicciones están fuertemente influenciadas por esta variable. Entre las categorías más importantes, se encuentra que ser Isapre y que el asegurado sea hombre generan un aumento significativo en el error.

4.2.8.2.3. Random Forest Reembolso

De acuerdo con lo planteado en 2.3.5, se entrenarán distintos modelos basados en las posibles combinaciones de variables y combinaciones de hiperparámetros, con el objetivo de encontrar la combinación óptima que entregue los mejores resultados. El conjunto de predictores puede encontrarse en la seccion B.6.1, mientras que los hiperparámetros en la tabla 59. A continuación, se presentan las métricas obtenidas en la tabla 60:

Tabla 59: Hiperparámetros Random Forest Reembolso

Parámetros	**Valor**
Árboles	500
Variables a considerar	12
Tamaño Mínimo Nodo	1
Variables por División	9

Tabla 60: Métricas de Desempeño del Modelo Ranger Severidad (Reembolso Aplicado)

Métrica	**Valor**
MAE (Mean Absolute Error)	0.032
MSE (Mean Squared Error)	0.0046
MAPE (Mean Absolute Percentage Error)	19.7357 %

En comparación con los modelos anteriores, se obtiene una mejora considerable, especialmente en términos del MAPE, donde se logra una mejora del 6 %.

Reales vs Predichos: Para evaluar el poder predictivo del modelo, se estudia su desempeño con la base de testeo (figura 41):

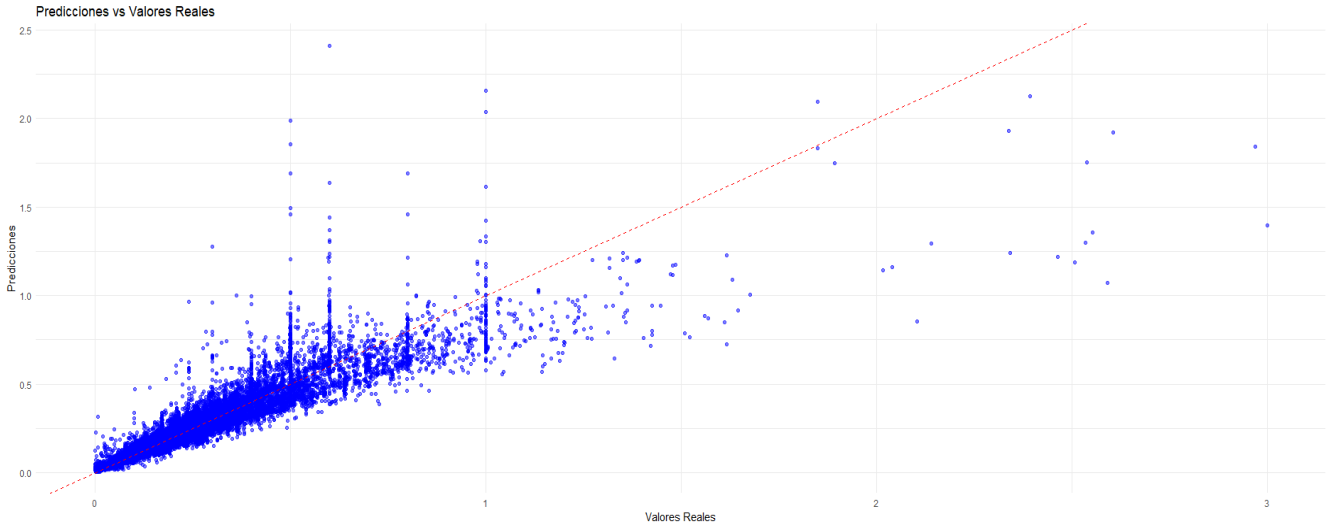

Figura 41: Gráficos Reales vs Predicciones Random Forest Reembolso

Se observa una mejora con los datos menores a una UF, donde la predicción muestra, a simple vista, una tasa de error baja. Lo importante es la existencia de valores con gran variabilidad en los resultados, lo cual puede deberse a la falta de predictores que expliquen estos resultados. Por lo tanto, los modelos tienen un margen de mejora a futuro.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera (figura 42):

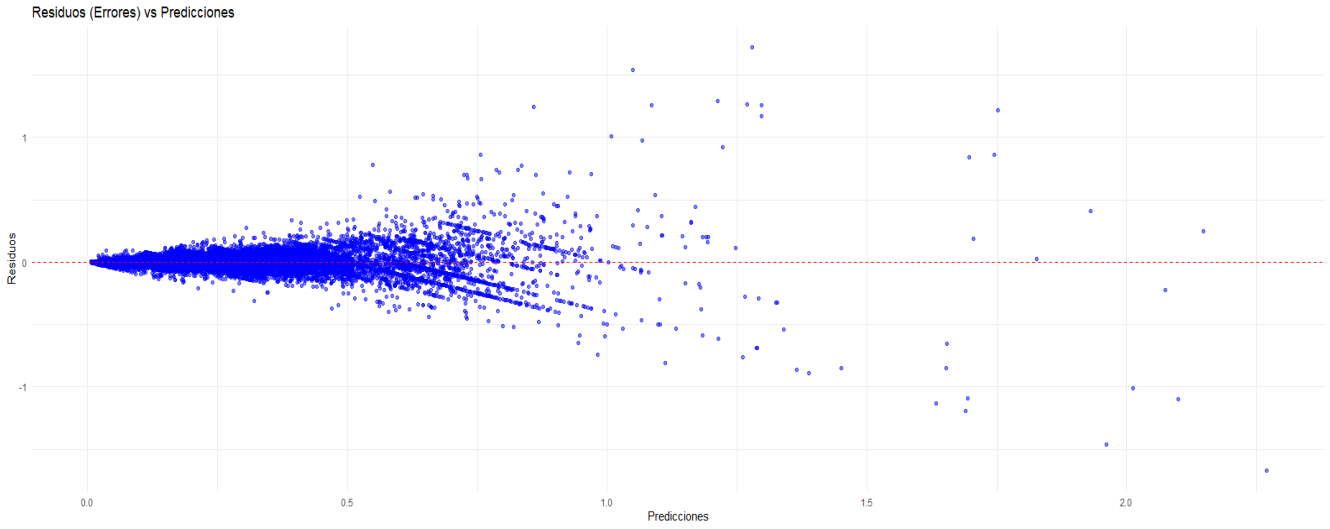

Figura 42: Gráficos de Errores Random Forest Reembolso

Aquí, se observa lo mismo que se mencionó anteriormente: los errores del modelo son heterocedásticos, lo que indica que al modelo tiene problemas para predecir valores altos de reembolso. Sin embargo, se observa que los datos outliers presentan una menor escala del error.

Importancia de Variables: Finalmente, a la hora de valorar las variables que más aportan al modelo, se pudo encontrar lo siguiente (figura 43):

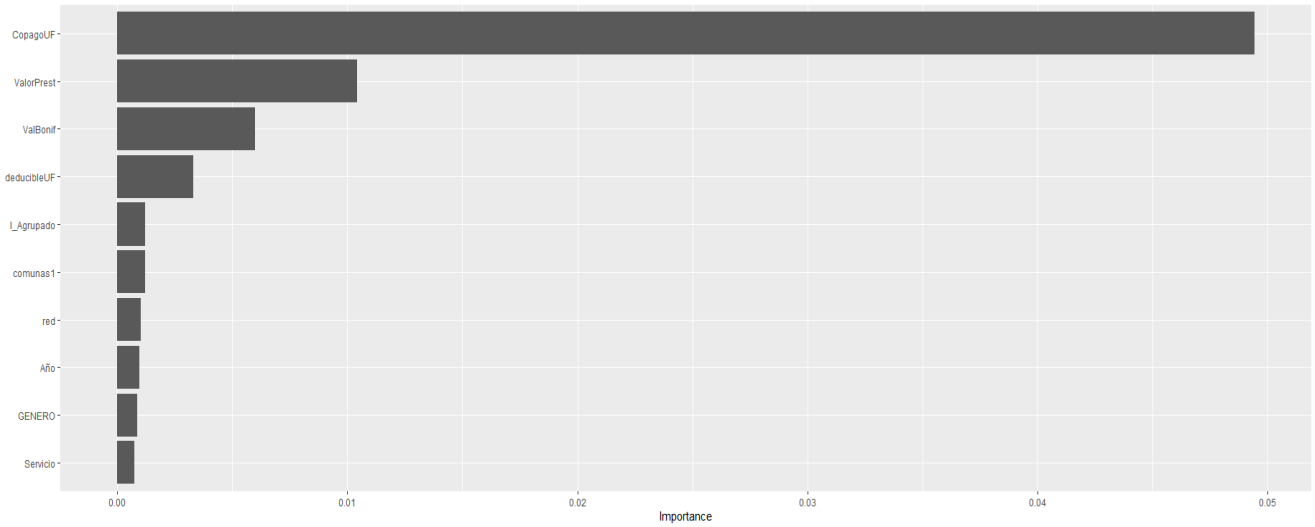

Figura 43: Gráficos de Importancia Random Forest Reembolso

Como se puede observar en la figura 43, la variable más importante para el modelo al determinar el reembolso aplicado es el copago. Esto implica que para el modelo es mucho más relevante contar con la información sobre el valor por el cual se aplicará el reembolso, en lugar de los valores por separado. Cabe mencionar que, aunque en el análisis descriptivo se encontró que el deducible no aportaba debido a la multicolinealidad, esta variable presenta una gran importancia en el modelo.

4.2.8.2.4. Boosted Trees Reembolso

Similar a lo planteado en el modelo anterior, se entrenarán múltiples modelos, donde el vector de predictores puede ser encontrado en B.6.2 y los hiperparámetros(tabla 61) que entrega los siguientes resultados resumidos en la tabla 62:

Tabla 61: Hiperparámetros Boosted Trees Reembolso

Parámetros	**Valor**
Objetivo	reg:gamma
Árboles	500
Variables a considerar	7
Tamaño Mínimo Nodo	1

Tabla 62: Métricas de Desempeño del Modelo Boosted Trees Reembolso

Métrica	**Valor**
MAE (Mean Absolute Error)	0.033
MSE (Mean Squared Error)	0.0045
MAPE (Mean Absolute Percentage Error)	16.90 %

Se logra encontrar una ligera mejora en comparación al modelo anterior, en términos de MAPE, pero en términos de error absoluto (MAE), hay un ligero aumento del error, lo que puede indicar que, aunque

globalmente el error es menor, este modelo presenta observaciones outliers con mayor error que el modelo anterior.

Reales vs Predichos: Para ver el poder predictivo del modelo, se estudia su desempeño con la base de testeo (figura 44):

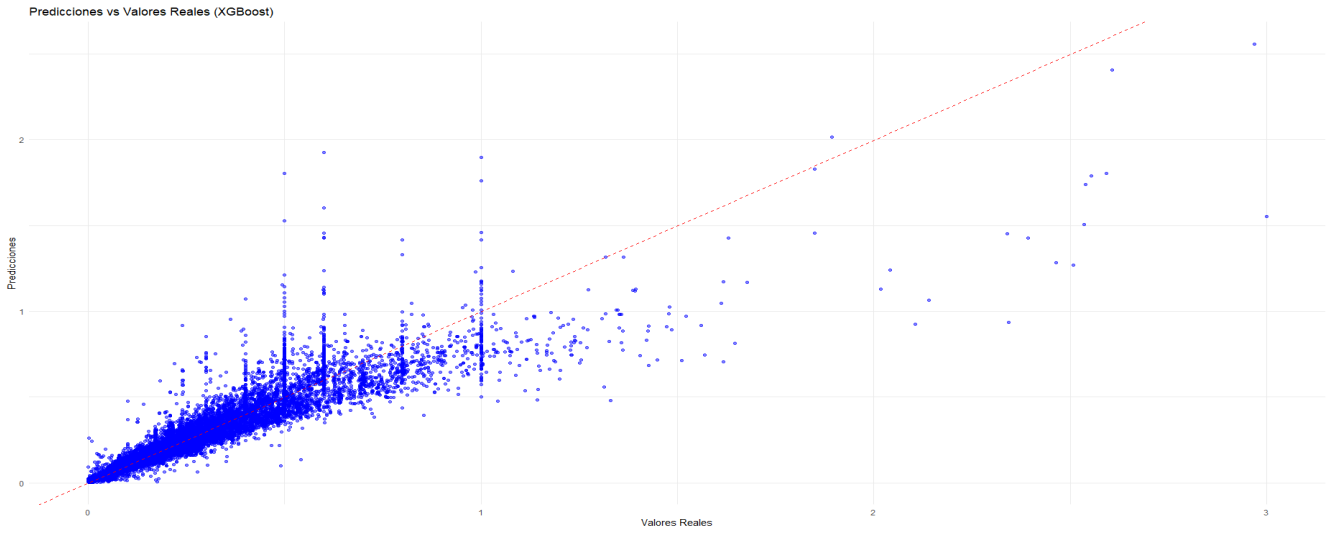

Figura 44: Gráficos Reales vs Predicciones Boosted Trees Reembolso Aplicado

Se observa que los resultados son similares a los propuestos con Random Forest, pero este modelo es capaz de realizar predicciones con menor error para los datos outliers del modelo.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera (figura 45):

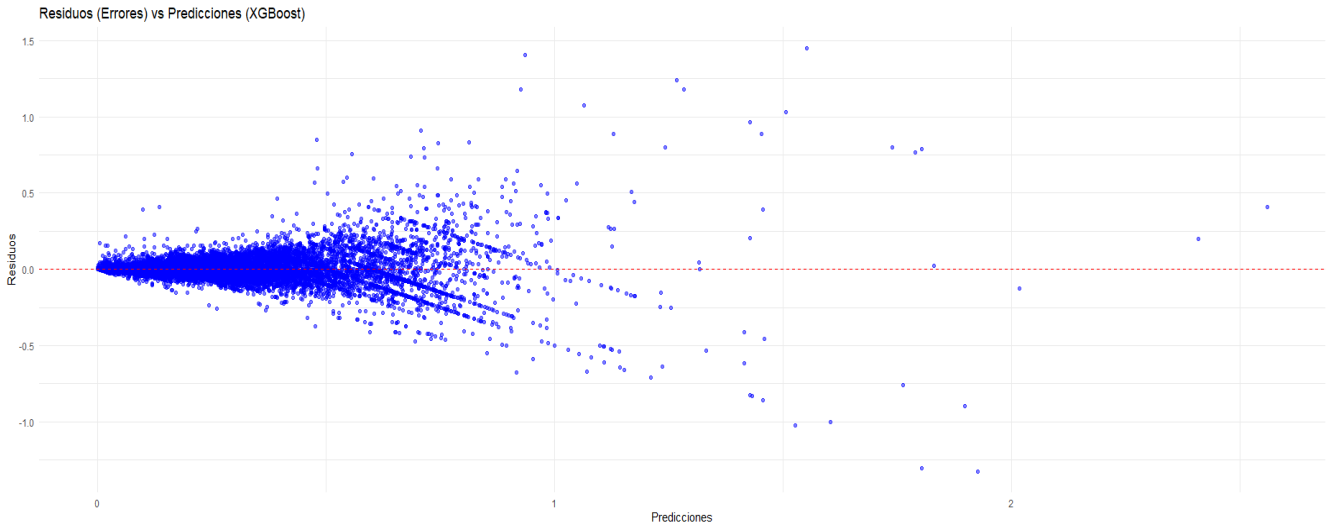

Figura 45: Gráficos de Errores Boosted Trees Reembolso Aplicado

Se puede observar que, conforme a lo mencionado, este modelo presenta errores con una escala mayor que el de Random Forest, pero al estudiar los outliers, se encuentra que la escala es mucho menor, y no que el dato atípico corresponde a una observación que no fue capaz de predecir de manera correcta.

Importancia de Variables: Finalmente, a la hora de valorar las variables que más aportan al modelo, se pudo encontrar lo siguiente(figura 46):

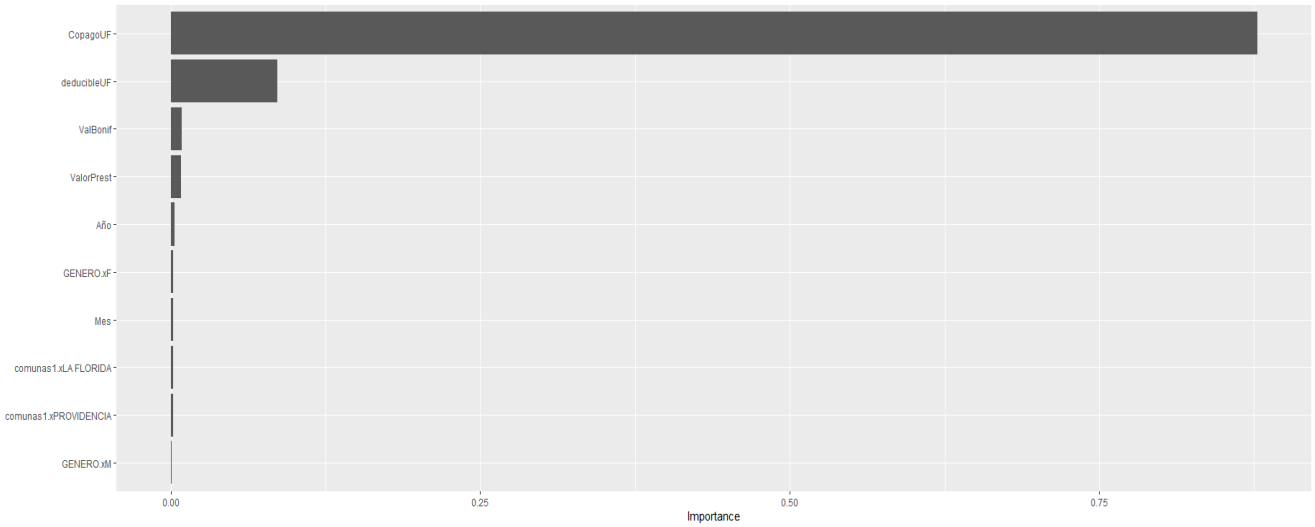

Figura 46: Gráficos de Importancia Boosted Trees Reembolso Aplicado

A diferencia del modelo anterior, la importancia del copago para las predicciones es mucho más predominante, siendo que el aporte del resto de variables es mucho más pequeño, lo cual plantea la discusión sobre el hecho de incluirlas.

4.2.9. Modelación del Número de Siniestros

En los resultados de la modelación del Número de siniestros, como se observó en el análisis descriptivo, es una variable de conteo con una sobredispersión clara, lo cual es una limitante, puesto que los modelos tradicionales no pueden lidiar con este problema. Se mostrarán los resultados con los modelos expuestos. Las variables utilizadas son las siguientes:

Tabla 63: Lista de variables categóricas y continuas para el modelo de Número de Siniestros.

Categóricas	**Continuas**
Año	DeducibleUF
Mes	BonificacionUF
Comunas	ValorPrestacionUF
Reg	CopagoUF
Relación	
Prestadores	
Primera capa(IAgrupado)	

4.2.9.1. Modelo de Regresión Poisson

Se utiliza un modelo estadístico como base con el objetivo de obtener variables interpretables y proporcionar información de interés para la compañía, como la tasa de ocurrencia de los eventos. Inicialmente, el modelo se ajustó con todas las variables, y posteriormente se redujo mediante el método stepwise (modelo

reducido puede encontrarse en B.6). A continuación, se presentan los coeficientes más relevantes del modelo (tabla 64):

Tabla 64: Resultados del Modelo Poisson(Mayores Impactos)

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Copagouf	0.2322	0.0098	23.636	$< 2 \times 10^{-16}$
RelTitular	0.4329	0.0076	57.177	$< 2 \times 10^{-16}$
I_Agrupado_Otros	-0.9728	0.0177	-54.867	$< 2 \times 10^{-16}$
GENERO_F	0.8646	0.0167	51.764	$< 2 \times 10^{-16}$
Servicio_General	0.6278	0.0121	51.719	$< 2 \times 10^{-16}$
GENERO_M	0.8591	0.0167	51.432	$< 2 \times 10^{-16}$

En la tabla 64, la variable con mayor peso es el copago, que, por cada unidad de aumento, incrementa la tasa en 0.23. Además, cabe destacar que ambos géneros tienen un gran impacto y que el servicio con más frecuencia es el general. También se observa que los asegurados fuera del sistema Isapre-Fonasa disminuyen la tasa de siniestros, por lo que tenerlos es más valioso para la compañía.

A continuación, se analizará el supuesto de equidispersión del modelo. Aunque, gráficamente, ya se ha observado que este supuesto no se cumple, es necesario determinar la tasa real para su estudio. Para ello, se empleará el estadístico planteado en A.4.4.

Tabla 65: Resultados del Test de Sobredispersión

Métrica	**Valor**
Estadístico de Sobredispersión	4.5914
Grados de Libertad Residuales	$n - p$
Interpretación	Existe evidencia de sobredispersión

Ya evaluado el supuesto se seguirá con la evaluación del modelo, del cual se obtuvieron las siguientes métricas (tabla 66):

Tabla 66: Métricas del Modelo Regresión Poisson

Métrica	**Valor**
MAE (Mean Absolute Error)	3.8074
MSE (Mean Squared Error)	3178.2600
MAPE (Mean Absolute Percentage Error)	127.0108 %
AIC (Criterio de Información de Akaike)	264785

De lo cual se puede observar que los valores son excesivamente altos, lo que indica que el modelo no esta explicando la tasa de ocurrencia de la variable respuesta.

Reales vs Predichos: Para ver el poder predictivo del modelo, se estudia su desempeño con la base de testeo (figura 47):

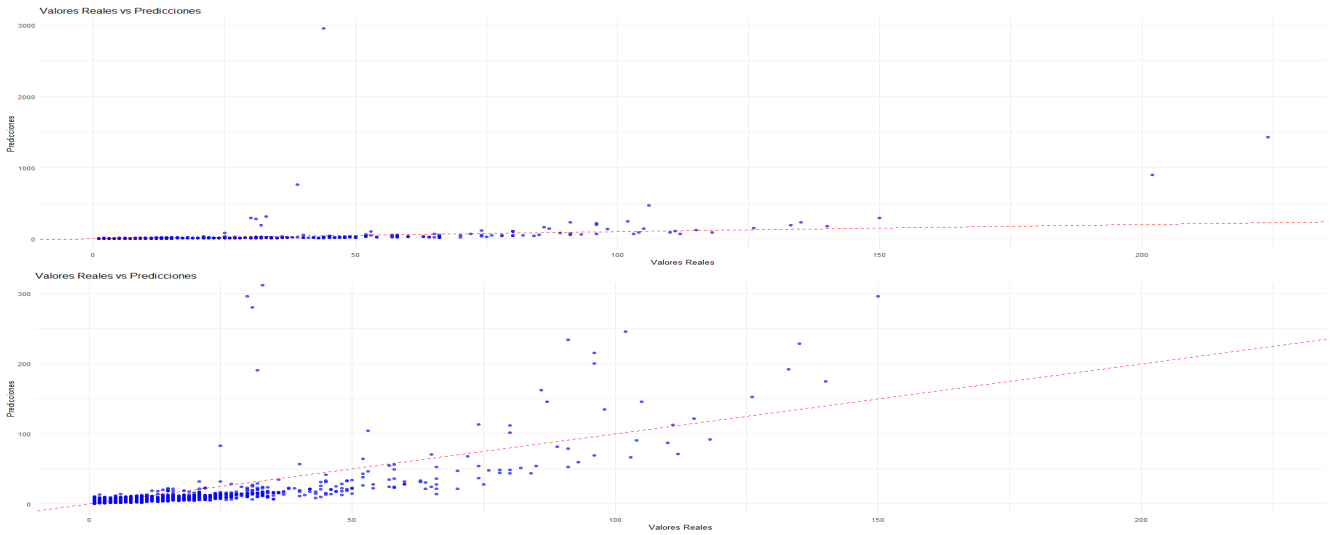

Figura 47: Gráfico Reales vs Predichos Regresión Poisson

Se observa que el modelo no tiene un buen ajuste, pero que los datos outlier no los logra predecir de manera correcta si no que cae en una sobre predicción muy severa afectando todo el rendimiento, si se observa el efecto especifico se observa que el modelo tampoco tiene un buen ajuste en valores bajos puesto que se observa que el modelo subestima los valores cercanos al cero.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera (figura 48):

Figura 48: Gráfico de Errores Regresión Poisson

Podemos observar que el modelo es altamente heterocedastico, puedo que el error del modelo aumenta linealmente en términos de los valores ajustados, a pesar de ser un modelo adecuado en teoría, el modelo poisson no entrega los resultados esperados y se espera que los modelos de machine learning mejoren lo entregado.

4.2.9.2. Redes Neuronales Número Siniestros

En este caso, se entrenó el modelo utilizando una pérdida de Poisson y probando múltiples arquitecturas que permitieran capturar las relaciones no lineales. La arquitectura final utilizada puede encontrarse en B.3 y los hiperparámetros utilizados, junto con las métricas de desempeño, se presentan en las tablas 67 y 68.

Tabla 67: Hiperparámetro Redes Neuronales Número Siniestros

Hiperparámetro	**Valor**
Dropout	0.2
Batch size	64
Learning rate	10^{-5}
$L2_rate$	10^{-3}

Tabla 68: Métricas del Modelo de Red Neuronal Número Siniestros

Métrica	**Valor**
MAE (Mean Absolute Error)	1.99
MSE (Mean Squared Error)	43.11
MAPE (Mean Absolute Percentage Error)	34.92 %

Donde podemos observar en la tabla 68 que las métricas son bastante altas, dando a entender que el modelo no fue capaz de capturar las relaciones existentes, otra posibilidad es que el modelo fue sobreentrenado y no fue capaz de predecir nuevas observaciones, afectando a las nuevas predicciones.

Reales vs Predichos: Para ver el poder predictivo del modelo, se estudia su desempeño con la base de testeo (figura 49):

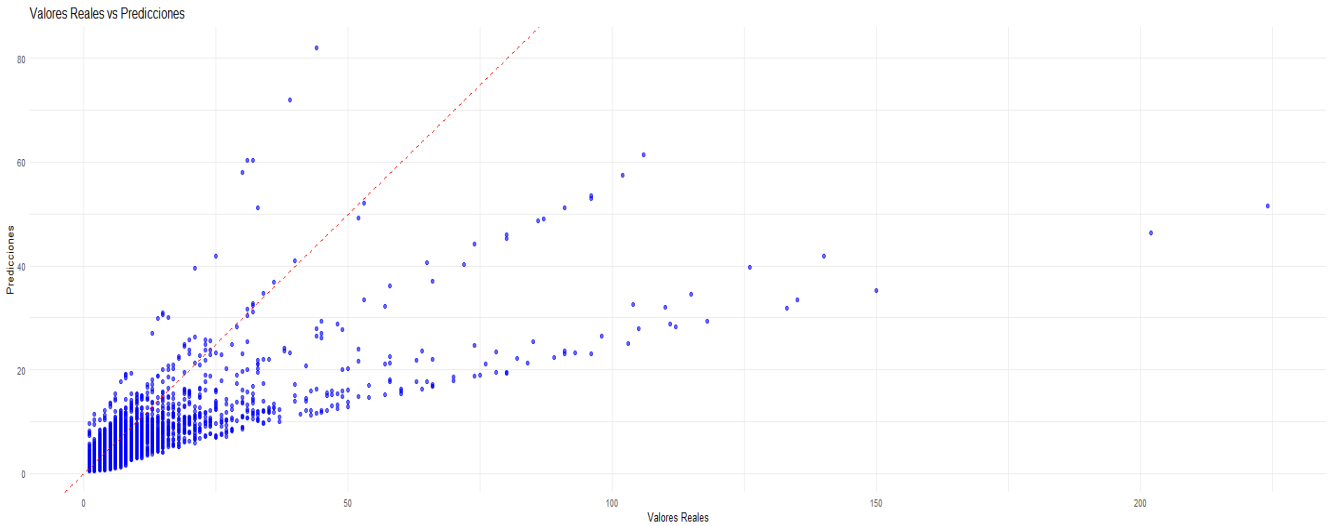

Figura 49: Gráfico Reales vs Predichos Red Neuronal Número Siniestros

Se observa que el modelo no fue capaz de capturar las relaciones, donde se observa que tiene una alta tendencia a subestimar los valores reales y que

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera (figura 50):

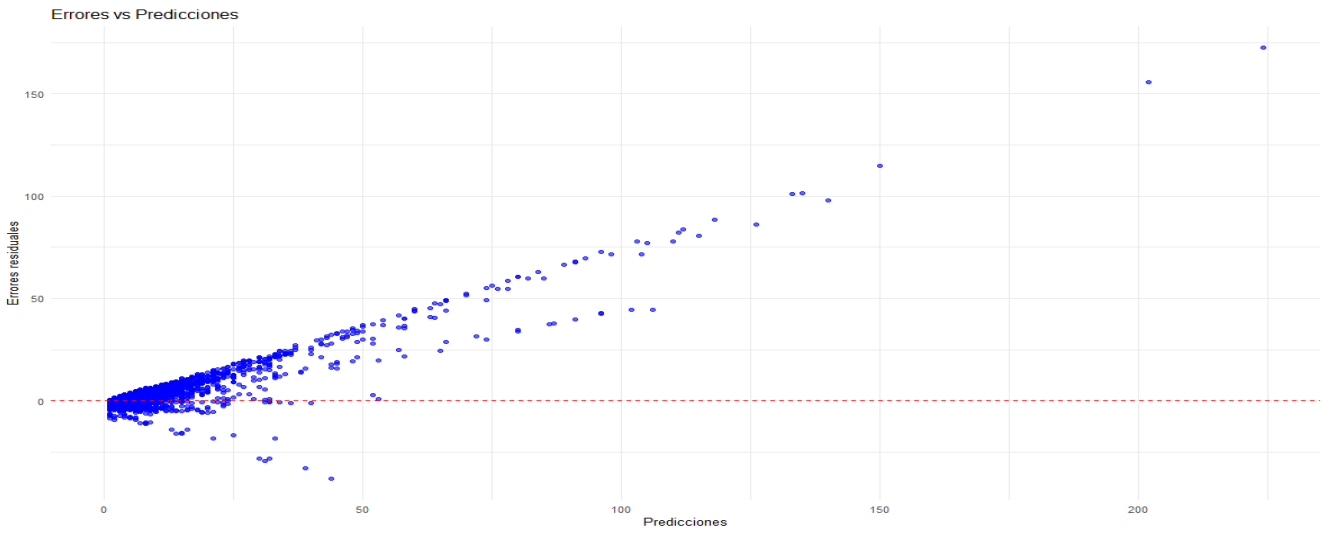

Figura 50: Gráficos de Errores Red Neuronal Número Siniestros

Se observa una alta heterocedasticidad en todo el modelo, donde en cada aumento del valor real, hay un aumento del error, por lo que el modelo no fue capaz de estimar los valores, y hay margen de mejora del resto de modelos.

Importancia de Variables: Finalmente, a la hora de valorar las variables que más aportan al modelo, se pudo encontrar lo siguiente (figura 51):

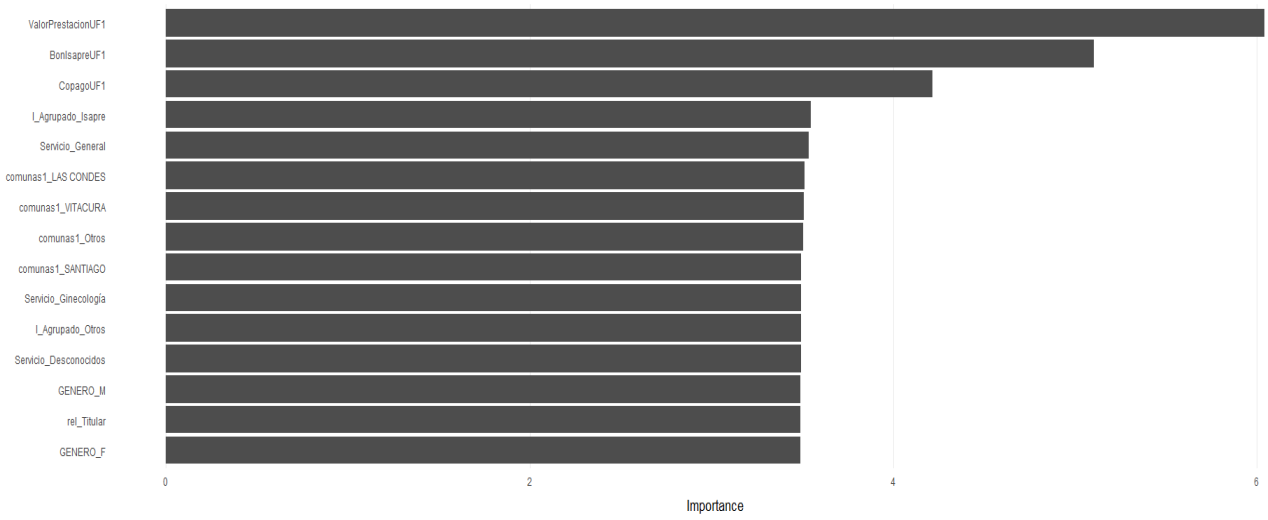

Figura 51: Gráficos de Importancia de Variables Redes Neuronales Número Siniestros

Donde se observa que la mayoría de las categorías tienen el mismo impacto y por lo tanto son intercambiables estar o no y las variables numéricas tienen un alto efecto en el error el cual no se ve reflejado en los resultados.

4.2.9.3. Random Forest Número Siniestros

En este caso, se utilizó la misma configuración del modelo que para modelar el reembolso, pero aplicada a la base de datos de siniestros, para lo cual se utilizó el siguiente vector de hiperparámetros(tabla 69) y obteniendo las siguientes métricas (tabla 70):

Tabla 69: Hiperparámetros Random Forest Número Siniestros

Parámetros	Valor
Árboles	500
Variables a considerar	11
Tamaño Mínimo Nodo	1
Variables por División	7

Tabla 70: Métricas del Modelo de Random Forest Número Siniestros

Métrica	Valor
MAE (Mean Absolute Error)	0.3926275
MSE (Mean Squared Error)	4.510793
MAPE (Mean Absolute Percentage Error)	7.245324 %

Las variables consideradas se pueden encontrar en la sección B.6.3. Se observa una gran mejora en comparación al modelo de redes neuronales, donde se espera que no tenga tantos problemas al determinar valores altos de frecuencia.

Reales vs Predichos: Para ver el poder predictivo del modelo, se estudia su desempeño con la base de testeo(figura 52):

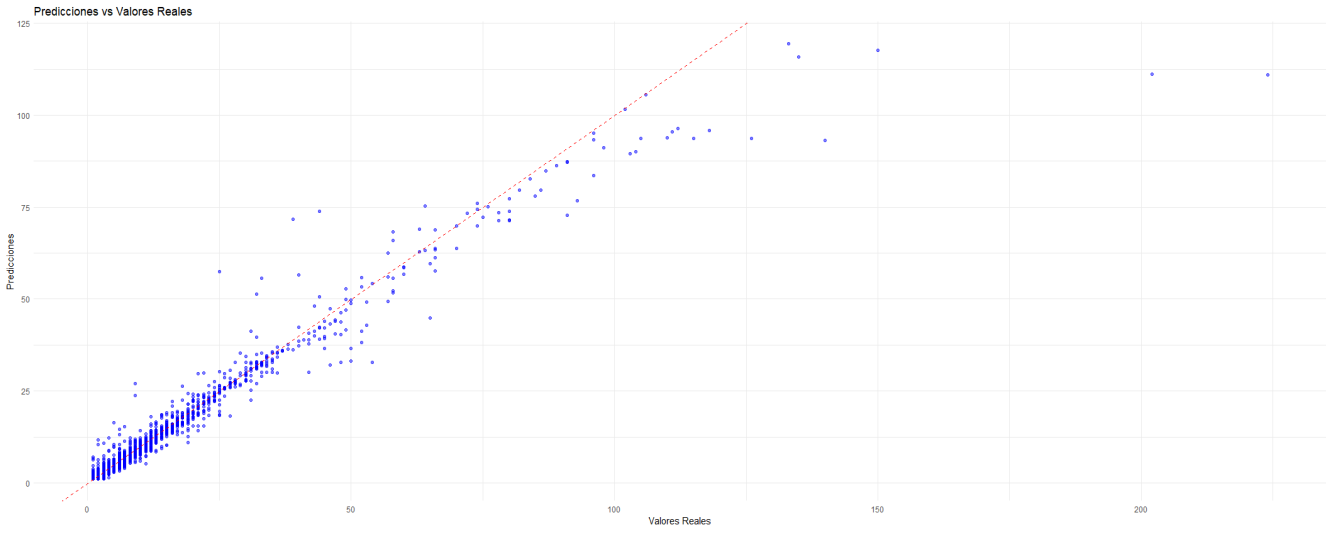

Figura 52: Gráficos de Reales vs Predichos Random Forest Número Siniestros

Podemos observar que se mantiene como buena opción, pero que en valores altos tiende a subestimar el valor real, lo cual se ve remarcado por 2 observaciones más altas, las cuales el modelo no es capaz de determinar el valor y lo subestima en gran medida. Lo anterior es un problema, puesto que una mala

estimación en este caso puede costar que una cuenta sea excesivamente más cara de lo que se espera. En cambio, con los valores bajos, se mantiene la tendencia anterior.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera (figura 53):

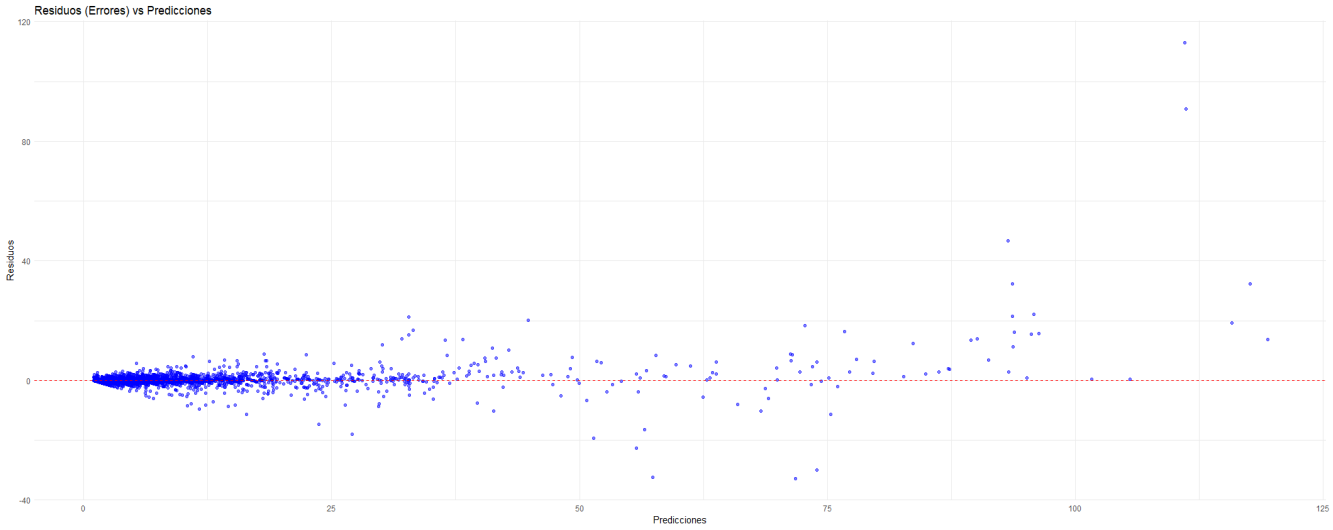

Figura 53: Gráficos de Errores Random Forest Número Siniestros

Observando los errores del modelo, a diferencia de los modelos anteriores se controla la subestimación. Sin embargo, hay 2 observaciones atípicas que, como se observó antes, el modelo no fue capaz de determinar de manera correcta su valor, lo que explica la subida de la medida del error, puesto que, como se observa, está controlado en la mayor parte de los valores.

Importancia de Variables: Finalmente, a la hora de valorar las variables que más aportan al modelo, se pudo encontrar lo siguiente (figura 54):

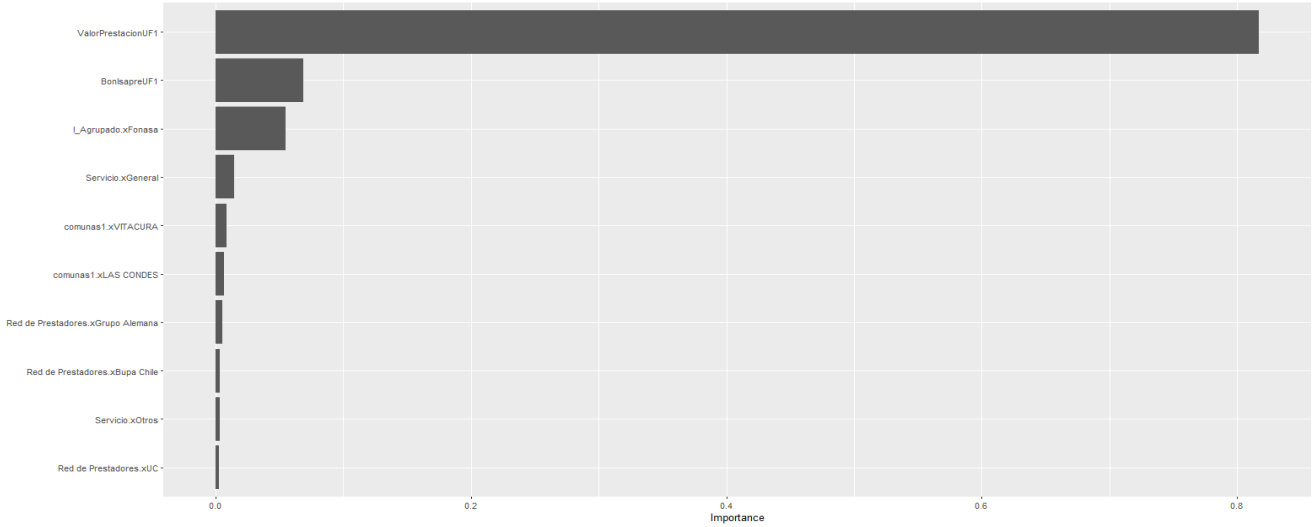

Figura 54: Gráficos de Importancia de Variables Random Forest Número Siniestros

Se observa en la figura 54 que la variable asociada al valor de la prestación es la que tiene mayor impacto en el modelo, lo que implica que la mayor parte de la varianza proviene de esta variable. Según este modelo,

esta variable es suficiente para explicar el fenómeno, tanto así que su ausencia implica que el MAE aumente a más del doble.

4.2.9.4. Boosted Trees Número Siniestros

En este caso, se utilizó la misma configuración del modelo anterior, para lo cual bajo los siguientes hiperparametros (tabla 71) y con los siguientes resultados(tabla 72):

Tabla 71: Hiperparámetros Boosted Trees Número Siniestros

Parámetros	**Valor**
Objetivo	reg:negativebinomial
Árboles	500
Variables a considerar	8
Tamaño Mínimo Nodo	5

Tabla 72: Métricas del Modelo Boosted Trees Número Siniestros

Métrica	**Valor**
MAE (Mean Absolute Error)	0.4005152
MSE (Mean Squared Error)	3.69579
MAPE (Mean Absolute Percentage Error)	8.321431 %

Las variables consideradas se pueden encontrar en la sección B.6.4. Se observa que, en comparación con el modelo anterior, hay un aumento tanto en el MAE como en el MAPE, lo que indica que, en promedio, hay un aumento del error. En cambio, se presenta una disminución del MSE, lo que indica que, a diferencia del modelo anterior, se espera que los errores obtenidos no tengan una magnitud de la misma escala que en Random Forest.

Reales vs Predichos: Para ver el poder predictivo del modelo, se estudia su desempeño con la base de testeo (tabla 55):

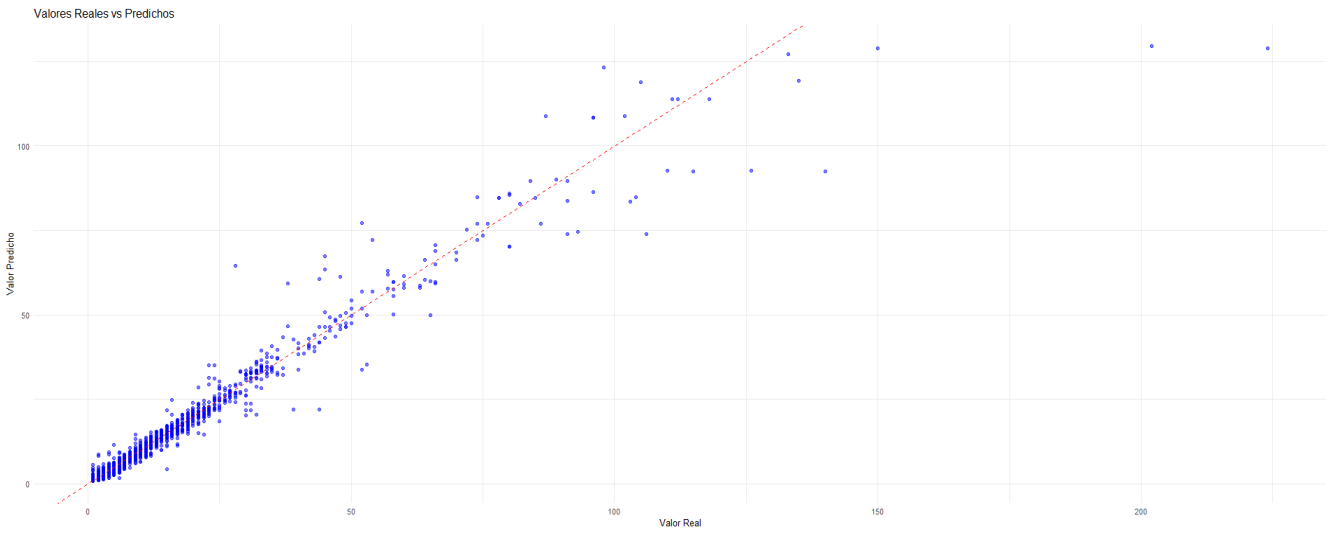

Figura 55: Gráficos de Reales vs Predichos Boosted Trees Número Siniestros

Podemos observar que se mantiene como buena opción, pero que en valores altos tiende a subestimar el valor real, lo cual se ve remarcado por dos observaciones más altas, las cuales el modelo no es capaz de determinar el valor y las subestima en gran medida. Lo anterior es un problema, puesto que una mala estimación en este caso puede costar que una cuenta sea excesivamente más cara de lo que se espera. En cambio, con los valores bajos, se mantiene la tendencia anterior.

Dispersión del Error: Se estudia el comportamiento del error, el cual se comporta de la siguiente manera (tabla 56):

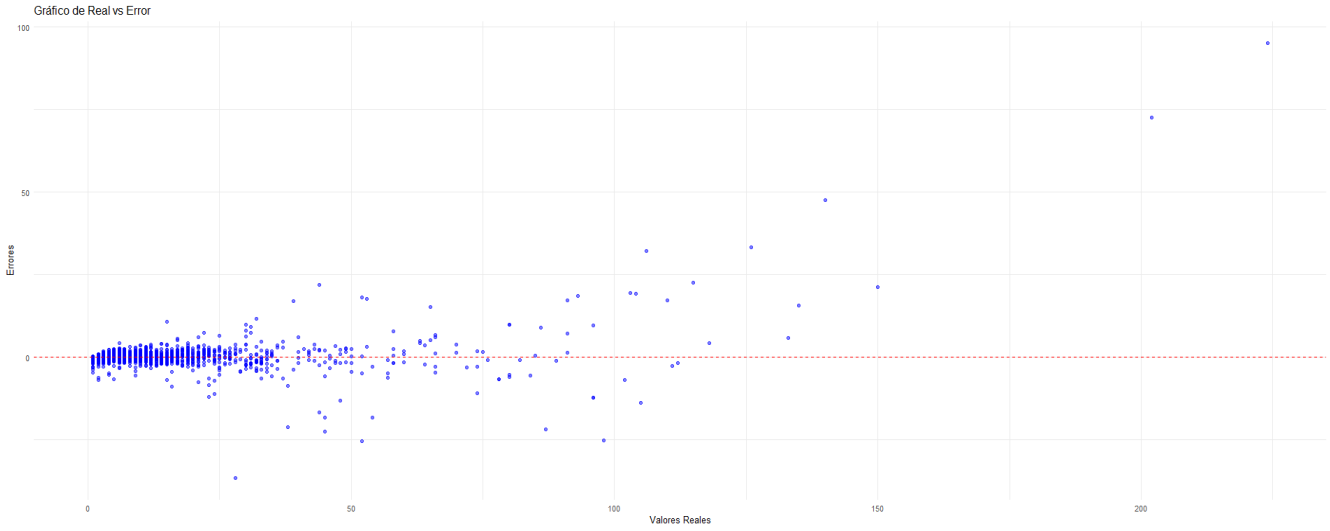

Figura 56: Gráficos de Errores Boosted Trees Número Siniestros

Se observa lo esperado, donde las observaciones presentan más dispersión en cuanto más aumenta la frecuencia. Dado esto, sigue siendo mejor opción el modelo anterior.

Importancia de Variables: Finalmente, a la hora de valorar las variables que más aportan al modelo, se pudo encontrar lo siguiente (tabla 57):

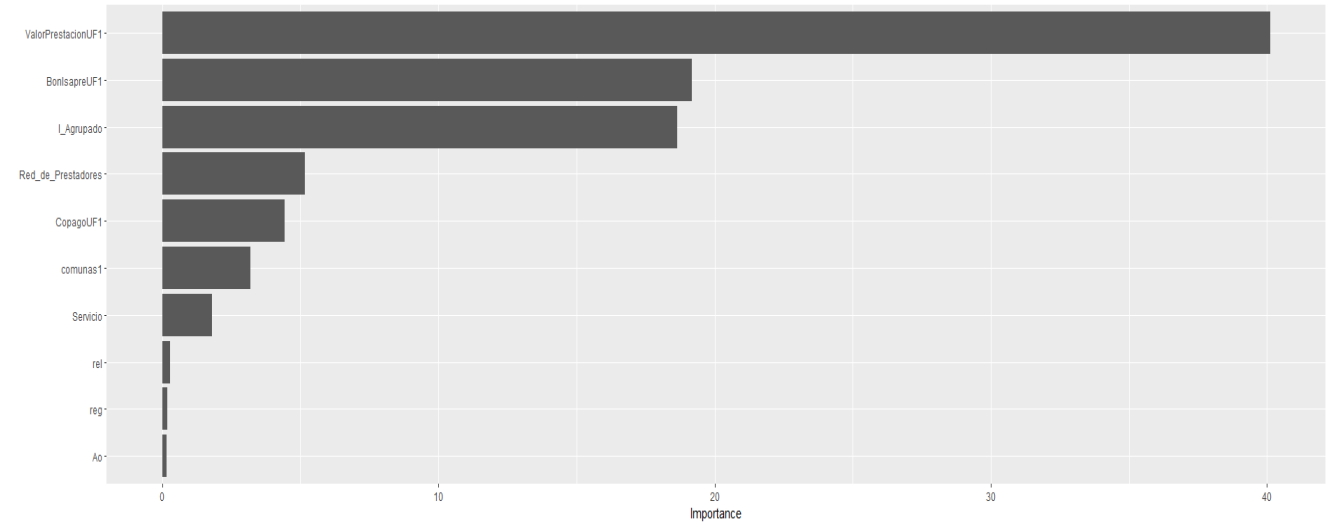

Figura 57: Gráfico de Importancia de Variables Boosted Trees Número Siniestros

Se observa en la figura 54 que las variables asociadas al valor de la prestación y la bonificación son las que tienen mayor peso en el modelo, lo que implica que es más interesante realizar estudios sobre el valor de la prestación y que los cambios en la bonificación de la primera capa tienen un mayor peso al momento de definir el reembolso. Cabe destacar que la red de prestadores y la primera capa del asegurado también tienen un peso importante al momento de realizar las predicciones.

4.2.10. Comparación de Modelos

En esta sección se compararán los resultados obtenidos de los distintos modelos.

4.2.10.1. Reembolso Aplicado

Etapa 1: En este caso, ambos modelos obtuvieron resultados similares, con un desempeño casi igual, siendo ligeramente superior el modelo de regresión logística. Sumado a la capacidad de entregar información de interés para la compañía, se elige utilizar este modelo.

Etapa 2: Se presenta un resumen de los modelos realizados:

Tabla 73: Comparación de las Métricas de Desempeño entre Modelos para el Reembolso Aplicado.

Métrica	**RML**	**Red Neuronal**	**Random Forest**	**Boosted Trees**
Mean Absolute Error	0.052	0.439	0.032	0.033
Mean Squared Error	0.094	0.0069	0.0046	0.0045
MAPE(%)	38.07 %	26.80 %	19.73 %	16.90 %

En este caso del reembolso (tabla 73), el modelo de regresión múltiple, debido a que no cumple con los supuestos, no es comparable. Sin embargo, los modelos restantes tienen un desempeño similar, destacando las redes neuronales, aunque manteniendo el problema de que los resultados no son satisfactorios, los cuales fueron mejorados por los modelos basados en arboles, del cual se destaca el modelo basado en bossted trees, el cual tiene un error global y cuadrático menor y que según los gráficos presenta menores problemas varianza del error.

4.2.10.2. Número de Siniestros

Tabla 74: Comparación de las Métricas de Desempeño entre Modelos de Número de Siniestros

Métrica	**Poisson**	**Red Neuronal**	**Random Forest**	**Boosted Trees**
Mean Absolute Error	3.8074	1.99	0.39	0.40
Mean Squared Error	3178.2600	43.11	4.51	3.69
MAPE (%)	127.0108 %	34.92 %	7.24 %	8.32 %
AIC	264785	No aplica	No aplica	No aplica

Dado los resultados obtenidos (table 74), el modelo basado en Random Forest fue el que obtuvo los mejores resultados, puesto que el modelo Poisson, a pesar de cumplir con el supuesto, no obtuvo buenos resultados. El modelo de redes neuronales, a pesar de tener resultados similares, tuvo un peor desempeño, incurriendo más en problemas de sobredispersión que el modelo Random Forest y el de boosted treed a pesar de tener resultados similares se observa que tiene menor error cuadrático pero mayor error medio, es

decir, que las mayores desviaciones tienen una menor magnitud pero goblamente sigue tiendo una tasa de error mayor.

4.2.11. Evaluación de Resultados

Finalmente, ya elegidos los modelos con mejor desempeño, es necesario calcular la frecuencia y la severidad con los resultados obtenidos anteriormente y evaluar la diferencia real con el objetivo de evaluar si los resultados no implican una pérdida monetaria importante. A continuación, se presentarán los valores calculados usando las fórmulas presentadas en 2.1.1:

Expuestos: Es necesario presentar el vector de expuestos usado (tabla 75), el cual corresponde a los titulares actuales de la compañía en el periodo de tiempo que se utilizó como testeo, es decir, año 2024 y meses (5,6,7,8):

Tabla 75: Valores Expuestos

Fecha	Valor
2024-05	24393
2024-06	25220
2024-07	29419
2024-08	29767

Resultados Frecuencia: A continuación se presentan los resultados finales para el cálculo de la frecuencia de consultas médicas ambulatorias:

Tabla 76: Resultados de Frecuencia

Año	Mes	Sin_Reales	Sin_Predichos	Frec_Real	Frec_Predicha	Diferencia
2024	5	10917.000	10717.000	0.448	0.439	**+3.19 %**
2024	6	8994.000	8983.000	0.357	0.356	**+0.31 %**
2024	7	10883.000	10702.000	0.370	0.364	**+1.69 %**
2024	8	11014.000	10793.000	0.370	0.363	**+1.89 %**

Donde podemos observar (tabla 76) que en la suma de los siniestros obtenidos mensualmente, la diferencia es ligera, donde en promedio, al calcular la frecuencia, la diferencia más grande corresponde a un +3.19 % del valor real, en cambio, la más baja es de un +0.31 %, lo que demuestra el poder de los modelos obtenidos. Además, con el error obtenido, implica que en un mes predicho con el procedimiento trabajado, la frecuencia de siniestros bajo nuestros expuestos es lo suficientemente baja como para que los valores no sean sobreajustados.

Resultados Severidad: A continuación se presentan los resultados finales para el cálculo de la Severidad de consultas médicas ambulatorias:

Tabla 77: Resultados Severidad

Año	Mes	Reemb_Real	Reemb_Predicho	Sev_Real	Sev_Predicha	Diferencia
2024	5	1742.000	1712.000	0.1596	0.1597	+0.08 %
2024	6	1497.000	1488.000	0.1665	0.1657	-0.48 %
2024	7	1759.000	1751.000	0.1616	0.1636	+1.24 %
2024	8	1925.000	1905.000	0.1748	0.1765	+1.01 %

Donde podemos observar(tabla 77) que los resultados son aún más cercanos a la realidad, con un error medio cercano al 0.4625 %, lo que, para la escala estudiada, implica que para un mes promedio predicho, la pérdida por siniestro está en la escala aproximada de 28 pesos chilenos.

Resultados Costo Medio: Finalmente, se calculan los valores predichos para el costo medio de consultas médicas ambulatorias:

Tabla 78: Resultados Costos Medios

Año	Mes	Costo Medio Real	Costo Medio Predicho	Diferencia Porcentual
2024	5	0.0714	0.0702	-1.68 %
2024	6	0.0594	0.0590	-0.66 %
2024	7	0.0598	0.0595	-0.52 %
2024	8	0.0647	0.0640	-1.08 %

Donde podemos observar que se presentan problemas de subestimar el valor real del costo medio, pero la magnitud de estos errores es muy cercana al 0 %, lo que implica que el error es imperceptible. Interpretando la diferencia, tenemos que para el mes 5 del año 2024, la diferencia en UF de lo que espera la compañía asegurar a alguien por consultas médicas ambulatorias es de un 0.0012 UF, que a día de hoy corresponde a 46.65 pesos de subestimación por persona.

Capítulo 5

Conclusiones

La presente investigación se enfocó en desarrollar una metodología para la modelación del costo medio mediante las variables que componen dicho índice (numero de siniestros y el reembolso), con el objetivo de aportar información relevante que permita mejorar los procesos de tarificación, como aportar con la proyección de los costos o creando indices basado en los coeficientes de los modelos.

El objetivo principal se cumplió al desarrollar una metodología que permite transformar la información de la cartera, complementándola con datos externos relevantes, como información geográfica, generando así una base de datos que facilita múltiples estudios y el análisis descriptivo de las variables. Asimismo, se realizó un análisis descriptivo identificando las variables más relacionadas con nuestro objeto de estudio, destacando especialmente que variables como comuna, servicio y primera capa presentan diferencias significativas entre sus categorías, además de observar las correlaciones establecidas con las variables de interés, donde se observo que presentan una alta correlación todas las variables utilizadas. Finalmente, se probaron múltiples modelos, utilizando modelos estadísticos tradicionales como base para medir la mejora obtenida al aplicar modelos de machine learning Los mejores resultados se obtuvieron con modelos basados en boosted trees para la predicción del reembolso, alcanzando un error medio (MAPE) de 16.9 %, mientras que para la predicción del número de siniestros, el modelo de Random Forest mostró el mejor desempeño con un error medio de 7.24 %.

En relación con trabajos previos, los resultados obtenidos en esta investigación difieren de los presentados por Tuininga (2022). Mientras que dicho estudio obtuvo resultados que beneficiaban el uso de modelos tradicionales (GLM) ante modelos de machine learning, debido a la sensibilidad que presentaban ante fluctuaciones en los precios, pero en nuestro estudio observamos lo contrario, logrando mejores resultados con métodos basados en machine learning, los cuales además mostraron una estabilidad adecuada expetuando por outliers especificos. Sin embargo, también hay oportunidades de mejora, ya que el estudio utilizó una gama más amplia de modelos, incluyendo métodos avanzados como Gradient Boosting y una mayor gama de enlaces para mejorar el ajuste de los modelos tradicionales.
Estas diferencias también se observan respecto al estudio de Guelman (2012), quien obtiene resultados más favorables para modelos GLM en comparación con modelos de boosted trees. Sin embargo, estas diferencias en el desempeño pueden explicarse debido a que su investigación se centra en seguros individuales, además de desconocerse el tratamiento exacto aplicado sobre las variables utilizadas.

Con quien se obtuvieron resultados similares fue conGraziadei et al. (2023), quienes reportaron un mejor desempeño al utilizar modelos de Random Forest en comparación con modelos Poisson para la modelación de la frecuencia de siniestros. Asimismo, los resultados encontrados coinciden con lo realizado por Cordeiro (2023), aunque desde un enfoque distinto, ya que su estudio modela la frecuencia y severidad para cierto grupo de clientes. En nuestro caso que se realizo en frente a un grupo de prestaciones, pese a lograr un desempeño similar, se obtuvo una tasa de error global ligeramente menor.

Como oportunidad de mejora futura, se plantea continuar evaluando otros modelos avanzados, como el Gradient Boosting propuesto por Guelman (2012), o modelos jerárquicos como los sugeridos por Zhang et

al. (2012). Asimismo, se recomienda explorar una mayor variedad de funciones de enlace en modelos GLM y técnicas adicionales como las Máquinas de Vectores de Soporte (SVM). Finalmente, sería conveniente considerar la incorporación de nuevas variables predictoras que logren explicar las relaciones que los modelos actuales no lograron capturar. Un ejemplo relevante es la variable edad, que, debido a limitaciones en su validación, no pudo ser incluida, pese a su importancia predictiva. Como limitación encontrada, se observó que los modelos basados en redes neuronales no fueron capaces de capturar las relaciones de la manera en que se esperaba, lo cual es consistente con lo reportado por Shwartz-Ziv y Armon (2021) y Grinsztajn et al. (2022), quienes señalan que los modelos basados en árboles tienen un mejor desempeño al tratar con datos tabulares. En particular, Grinsztajn et al. (2022) plantea que los modelos de redes son menos robustos al tratar con un mayor número de variables con poco aporte de información. Asimismo, otra explicación se encuentra en lo señalado por Werner y Modlin (2016), quien explica que cambios externos al seguro pueden impactar en los costos, como lo son las variaciones en la conciencia de reclamaciones, las prácticas judiciales u otros factores no económicos que no se pueden capturar ni modelar directamente.

Como trabajo futuro en otros proyectos, se propone aplicar y replicar esta metodología para el resto de los grupos de prestaciones, con el objetivo de obtener índices para todos los grupos y lograr proyecciones más precisas de los costos globales de la compañía. Otro trabajo futuro consiste en aprovechar los modelos planteados y realizar estudios de simulación con el objetivo de poder aprovechar el trabajo actual y proyectar valores más a futuro y casos hipotéticos de interés, como cambios en alguna ley o modificaciones en ciertos productos.

Anexo A

Anexo A: Información Adicional

A.1. Teoría de Credibilidad

A continuación se describirá una metodología clásica para la resolución de este problema. Dada una cartera de seguros, según Rincón (2012), en la teoría de credibilidad se estudian métodos de calculo de la prima pura, que combinan la experiencia individual (historial de reclamaciones) con la experiencia de grupo (comportamiento teórico). Que para el caso del estudio la experiencia individual corresponde al un seguro colectivo.

A.1.1. Credibilidad Total

Tenemos credibilidad total cuando se logran las condiciones tal que, se puede considerar que la experiencia individual es suficiente para calcular la prima pura.
Entonces sea, $S = s_1, ..., s_n$ el vector de reclamaciones (vector con información de los pagos realizados) para n pólizas y $\bar{S} = \frac{\sum_{i=1}^{n} s_i}{n}$, el gasto promedio por póliza, entonces se dice que S tienen credibilidad completa si:

$$\mathbb{P}(\|\bar{S} - \mathbb{E}(S)\| \leq k\mathbb{E}(S)) \geq p$$

Que sucede cuando la diferencia entre $\bar{S}$ y $\mathbb{E}(S)$ es muy pequeña, es decir, cuando $\bar{S}$ es un buen estimador de $\mathbb{E}(S)$. Típicamente se toman valores de $k = 0.05$ y $p = 0,9$ y bajo estas condiciones se considera que la prima pura es igual a $\bar{S}$.

A.1.2. Credibilidad Parcial

Para este caso, no se considera suficiente la experiencia personal para poder determinar la prima pura, entonces se toma como estimador de $\mathbb{E}(S)$ la siguiente combinación lineal, que combina la experiencia del asegurado con la del colectivo, entonces la prima pura viene dada por:

$$Prima = Z(n)\bar{S} + [1 - Z(n)]\mathbb{E}(S)$$

Donde $Z(n) \in [0 : 1]$ es el factor de credibilidad, este factor determina el peso que se le da a la experiencia, tanto individual como colectiva. Cuando $Z(n)$ es igual a 1 se supone credibilidad total. Ahora el problema es determinar el valor de $Z(n)$, el cual debe se obtenido ajustando la siguiente inecuación:

$$\mathbb{P}(Z(n)\|\bar{S} - \mathbb{E}(S)\| \leq k\mathbb{E}(S)) \geq p$$

A.1.3. Modelo de Bühlmann

El modelo de Bühlmann es un modelo proveniente de la teoría de credibilidad enfocado en estimar de forma lineal la prima correspondiente a un conjunto de asegurados que conforman una póliza. Este modelo tiene como característica ser de distribución libre, es decir, que no supone un distribución de los montos reclamados.
Sea X_{ij} la variable aleatoria que representa la experiencia de reclamaciones de la póliza i en el periodo j, entonces la mejor prima lineal dependiente de los datos observados esta dado por:

$$H[\mu(\theta_j)|X_{j1},...,X_{jn}] = Z(n)\bar{X} + [1 - Z(n)]m \qquad Z(n) = \frac{an}{an + s^2}$$

Donde:

- Parámetro de riesgo: θ_j corresponde al, el parámetro de la función de distribución del vector de reclamaciones de la póliza j. Este parámetro tiene las siguientes propiedades.
  - Homogénea en el tiempo: $\theta_{ij} = \theta_j$ para todo periodo $i = 1, ..., n$, es decir, todos los riesgos de la columna j tienen la misma distribución de probabilidad.
  - El vector del riesgo $\theta = (\theta_1, ..., \theta_k)$ son variables aleatorias independientes y con la misma distribución de probabilidad $\pi(\theta)$
- Prima pura Individual: $\mu(\theta_j) = \mathbb{E}(X_{ij}|\theta_j)$
- Prima pura colectiva: $m = \mathbb{E}[\mu(\theta_j)]$
- Indicador de la heterogeneidad de la cartera: $a = V[\mu(\theta_j)]$
- Medida global de la dispersión de la cartera: $s^2 = \mathbb{E}[V(X|\theta)]$

A.2. Algoritmo Random Forest

Sea $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ el conjunto de datos de entrenamiento, donde $x_i = (x_{i,1}, \ldots, x_{i,p})^T$. Para $j = 1$ hasta J:

1. Tomar una muestra bootstrap $\mathcal{D}_j$ de tamaño N de $\mathcal{D}$.
2. Usando la muestra bootstrap $\mathcal{D}_j$ se ajustar un árbol como modelo base.
   a) Comenzar con todas las observaciones en un solo nodo.
   b) Repetir los siguientes pasos recursivamente para cada nodo no dividido hasta que se cumpla el criterio de parada:
      i. Seleccionar una muestra aleatoria de m predictores de los p disponibles.
      ii. Encontrar la mejor división binaria entre todas las posibles para los m predictores seleccionados en el paso i.
      iii. Dividir el nodo en dos nodos descendientes usando la mejor división encontrada en el paso ii.

Para realizar una predicción en un nuevo punto x:

- **Regresión:**

$$\hat{f}(x) = \frac{1}{J}\sum_{j=1}^{J} \hat{h}_j(x)$$

- **Clasificación:**

$$\hat{f}(x) = \arg\max_y \sum_{j=1}^{J} I(\hat{h}_j(x) = y)$$

donde $\hat{h}_j(x)$ es la predicción de la variable respuesta en x usando el árbol j.

A.3. Test de Esfericidad de Bartlett

Bartlett propone verificar si el conjunto de variables en estudio presentan correlación mediante el estudio de la matriz de correlaciones poblacional R, para lo cual se propone la siguiente hipótesis a probar:

$$H_0 : |R| = 1 \quad v/s \quad H_1 : |R| \neq 1$$

La cual indica que bajo H_0 el determinante de R es igual a 1 y por lo tanto las variables son independientes entre sí. Esta dócima se comprueba mediante el siguiente estadístico de prueba:

$$T = -\left(n - 1 - \frac{2p+5}{6}\right) \ln |\hat{R}| \sim \chi^2_{p(p-1)/2}$$

Donde $|\hat{R}|$ es la matriz residual de correlaciones. Cuya región crítica está dada por:

$$R_\alpha = \{T > \chi^2_{1-\alpha,p(p-1)/2}\}$$

A.4. Indicador KMO

Con el mismo objetivo que el test de esfericidad, este es un indicador que evalúa la asociación entre pares de variables, condicionándolas a las variables restantes. Este indice relaciona la correlación observada entre los pares de variables con sus parciales de la siguiente manera:

$$KM0 = \frac{\sum_{i=1}^{p}\sum_{j\neq 1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j\neq 1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j\neq 1}^{p} \phi_{ij}^2}$$

Donde r_{ij} corresponde a la correlación de Pearson muestra entre las variables i y j, y ϕ_{ij} es la correlación parcial. Cuando este indice se aproxima al 0, se hace menos necesario realizar la descomposición mediante componentes principales, pero se aplicaran criterios segun el valor del indice:

Tabla 1: Tabla Valores KMO

Valores	**Estrategia**
$KMO > 0.75$	Indica fuerte dependencia entre variables, viabilidad del CP
$0.5 \leq KMO \leq 0.75$	Indica dependencia moderada entre las variables , no se descarta el CP
$KMO < 0.5$	Indica posible ausencia de correlación entre variables, inviable del CP

A.4.1. Indicador MSA

Indice construido a partir de KMO cuando este toma un valor muy bajo, permite identificar cual par de variables esta ocasionando este problema, este esta dado por:

$$MSA_i = \frac{\sum_{j\neq 1}^{p} r_{ij}^2}{\sum_{j\neq 1}^{p} r_{ij}^2 + \sum_{j\neq 1}^{p} \phi_{ij}^2}$$

Donde cuando si el MSA de la variable i es cercano a cero indica que existe una dependencia debil o inexistente con el resto de variables.

Análisis de Clasificación

A.4.2. Análisis de Sensibilidad y Especificad

Dado el mejor modelo, las probabilidades estimadas pueden utilizarse para asignar o clasificar el atributo de interés. La efectividad de esta clasificación depende en gran medida de la elección del umbral adecuado. Este umbral define el punto a partir del cual se considera que un evento ocurre o no, afectando directamente el balance entre sensibilidad y especificidad, y por ende, la calidad general de las predicciones.
Considerando p_i a la probabilidad modelada para la observación i se realiza la siguiente clasificación:

$$\hat{Y} = \begin{cases} 1 \; p_i > \text{umbral} \\ 0 \; p_i \leq \text{umbral} \end{cases}$$

Utilizando lo anterior es posible construir una tabla que clasifica los valores observados de Y con los predichos por el modelo ajustado $\hat{Y}$:

Tabla 2: Análisis de Sensibilidad

	Valores Predichos		
Valores Reales	Negativo $\hat{Y}=0$	Positivo $\hat{Y}=1$	Totales
Negativo $Y=0$	Verdadero Negativo (VN)	Falso Positivo (FP)	VN+FP
Positivo $Y=1$	Falso Negativo (FN)	Verdadero Positivo (VP)	FN+VP

- **VP**: Número de observaciones positivas clasificadas correctamente como positivas.
- **FN**: Número de observaciones positivas clasificadas incorrectamente como negativas.
- **FP**: Número de observaciones negativas clasificadas incorrectamente como positivas.
- **VN**: Número de observaciones negativas clasificadas correctamente como negativas.

A partir de estas clasificaciones se construyen los siguientes indicadores:

- **Sensibilidad o Precision:** Corresponde a la probabilidad de clasificar correctamente a la observación cuando esta presenta el atributo en estudio ($Y = 1$).

$$\mathbb{P}(\hat{Y} = 1 | Y = 1) = \frac{VP}{VP + FN}$$

- **Especificidad:** Corresponde a la probabilidad de clasificar correctamente a la observación cuando esta no presenta el atributo en estudio ($Y = 0$).

$$\mathbb{P}(\hat{Y} = 0|Y = 0) = \frac{VN}{VN + FP}$$

- **Exactitud:** Proporcion de casos clasificados correctamente:

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN}$$

A.4.3. Curva ROC

Estas curvas **relacionan la sensibilidad en función de los falsos positivos** (1-especificidad) y se genera a partir de la unión de todos los posibles umbrales de corte, esto conduce a que se pueda determinar un punto de corte óptimo. La curva ROC mide la capacidad del modelo de discriminar una variable dicotómica a partir del índice AUC (Área bajo la curva), de modo que mientras mayor sea el AUC, mayor es el poder de discriminación del modelo.

Valor AUC	Capacidad de Discriminar
$0.5 \leq AUC < 0.7$	Baja
$0.7 \leq AUC < 0.9$	Moderada
$0.9 \leq AUC \leq 1$	Alta

A.4.4. Supuestos Regresion Poisson

Supuestos

- Independencia: Las observaciones deben ser independientes entre sí, pudiéndose estudiar con pruebas de aleatoriedad.
- Distribución de Y: La regresión se basa en el supuesto de que la variable respuesta sigue una distribución de Poisson, lo que se puede estudiar mediante pruebas de bondad de ajuste bajo la siguiente hipótesis:

$$H_0 : Y \sim \text{Poisson}(\lambda);; \text{vs};; H_1 : Y \not\sim \text{Poisson}(\lambda) \tag{1}$$

- Equisdispersión: Se debe cumplir que la varianza de la variable respuesta sea igual a su esperanza por propiedad de la distribución de Poisson. En caso de no cumplirse, se entraría en los siguientes escenarios:
  - Sobredispersión: Sucede cuando $\mathbf{V}(Y) > \mathbf{E}(Y)$, lo que implica sobrestimar la varianza de los
  - Subdispersión: Sucede cuando $\mathbf{V}(Y) < \mathbf{E}(Y)$, lo que implica subestimar la varianza de los

  Nota: El supuesto de equisdispersión es suficiente para justificar el uso del modelo en vez de otros.

En el caso de que la información disponible presente significativamente una sobre o subdispersión, el uso de esta distribución como modelo queda descartado, ya que la sobredispersión lleva a subestimar los

errores estándar, lo que conduce a inferencias sesgadas acerca de la información obtenida del modelo. Por lo que es de suma importancia verificar este supuesto, de modo que se tiene lo siguiente:

$$H_0 : \mathbb{V}(Y) = \mathbb{E}(Y) \text{ vs } H_1 : \mathbb{V}(Y) > \mathbb{E}(Y)$$

Estadístico

$$RV = -2\left[\hat{L}(\text{Poisson}) - \hat{L}(\text{Binomial Negativa})\right] \sim \chi^2_1$$

Región crítica

$$RC = \{\text{m.a.}(n) \mid RV > \chi^2_{(1;1-\alpha)}\}$$

Anexo B

Anexo B: Modelos

B.1. Modelo Binario

```
modelo_keras <- keras_model_sequential() %>%
  layer_dense(units = 256, activation = "relu", input_shape = ncol(x_train),
              kernel_regularizer = regularizer_l2(0.0001)) %>%
  layer_batch_normalization() %>%
  layer_dense(units = 128, activation = "gelu") %>%
  layer_dense(units = 64, activation = "gelu") %>%
  layer_dense(units = 32, activation = "gelu") %>%
  layer_dense(units = 1, activation = "sigmoid")  # Activación sigmoid para clasificación binaria

# Compilar el modelo con función de pérdida y métricas ajustadas
modelo_keras %>% compile(
  loss = "binary_crossentropy",  # Pérdida para clasificación binaria
  optimizer = optimizer_nadam(learning_rate = 0.01),  # Optimizador Nadam
  metrics = c("accuracy", metric_auc())  # Métricas de clasificación
)
```

B.2. Modelo Redes Neuronales Lineal

```
  model <- keras_model_sequential() %>%
    # Capa entrada con regularización mejorada
    layer_dense(units = 1024, activation = "gelu", input_shape = ncol(x_train),
                kernel_regularizer = regularizer_l2(0.01)) %>%
    layer_batch_normalization() %>%
    layer_dropout(rate = params$dropout_1) %>%
    layer_dense(units = 512, activation = "relu") %>%
    layer_dense(units = 256, activation = "gelu") %>%
    layer_dense(units = 128, activation = "gelu") %>%
    layer_dense(units = 64, activation = "relu") %>%
    layer_dense(units = 32, activation = "gelu") %>%
    layer_dense(units = 1, activation = "linear")

  # Optimización del compilador
  model %>% compile(
```

```
    loss = loss_huber(),
    optimizer = optimizer_adam(
      learning_rate = params$learning_rate
    ),
    metrics = c("mape")
  )
```

B.3. Modelo Redes Neuronales Conteo

```
modelo_keras_conteo <- keras_model_sequential() %>%
    layer_dense(
      units = 128, activation = "relu",
      kernel_regularizer = regularizer_l2(params$l2_rate)
    ) %>%
    layer_dense(
      units = 64, activation = "relu",
      kernel_regularizer = regularizer_l2(params$l2_rate)
    ) %>%
    layer_batch_normalization() %>%
    layer_dropout(rate = params$dropout_rate) %>%
    layer_dense(
      units = 32, activation = "relu",
      kernel_regularizer = regularizer_l2(params$l2_rate)
    ) %>%
    layer_dense(units = 16, activation = "relu") %>%
    # Capa de salida con activación softplus => salida siempre > 0
    layer_dense(units = 1, activation = "softplus")

# Compilar modelo
modelo_keras_conteo %>% compile(
  loss = loss_poisson(),
  optimizer = optimizer_nadam(learning_rate = 0.001),
  metrics = c("mae", "mse","mape")
)
```

B.4. Resultados del Modelo Logístico

Tabla 1: Resultados del Modelo Logístico Reducido (Incluye Odds Ratio)

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
(Intercept)	1.95783	0.09915	19.747	<2e-16	7.08
deducibleUF	-29.76592	0.14923	-199.457	<2e-16	1.16e-13
CopagoUF	8.76669	0.13956	62.816	<2e-16	6.41e3
ValorPrest	-0.48882	0.13305	-3.674	0.000239	0.61
ValBonif	0.61582	0.13775	4.471	7.80e-06	1.85
Año_2020	-0.19892	0.04813	-4.133	3.58e-05	0.82
Año_2021	0.25138	0.05008	5.020	5.17e-07	1.29
Año_2022	-0.22489	0.02158	-10.419	<2e-16	0.80
Año_2024	0.19148	0.02591	7.391	1.46e-13	1.21
Mes_2	-0.05964	0.03417	-1.746	0.080865	0.94
Mes_3	-0.22835	0.02832	-8.063	7.47e-16	0.80
Mes_6	-0.33655	0.03437	-9.791	<2e-16	0.71
Mes_7	-0.17439	0.03514	-4.963	6.95e-07	0.84
Mes_9	-0.11110	0.03507	-3.168	0.001533	0.89
Mes_11	-0.04725	0.03233	-1.462	0.143847	0.95
Mes_12	0.05618	0.03302	1.702	0.088841	1.06
comunas1_LA.FLORIDA	0.71353	0.03699	19.289	<2e-16	2.04
comunas1_LAS.CONDES	0.14197	0.04422	3.211	0.001324	1.15
comunas1_MAIPU	0.26377	0.08465	3.116	0.001834	1.30
comunas1_MELIPILLA	0.26295	0.11175	2.353	0.018618	1.30
comunas1_Otros	-0.11400	0.05691	-2.003	0.045181	0.89
comunas1_RECOLETA	-0.13561	0.04592	-2.953	0.003147	0.87
comunas1_SANTIAGO	0.10872	0.04683	2.321	0.020265	1.11
reg_XIII.REGION.METROPOLITANA	-0.26681	0.05644	-4.728	2.27e-06	0.77
rel_Hijo.a	-0.34921	0.03207	-10.890	<2e-16	0.71
rel_Titular	-0.40490	0.02837	-14.274	<2e-16	0.67
GENERO_F	-0.90864	0.06801	-13.359	<2e-16	0.40
GENERO_M	-0.61390	0.06806	-9.020	<2e-16	0.54
I_Agrupado_Isapre	0.19199	0.02402	7.992	1.33e-15	1.21
I_Agrupado_Otros	-0.79567	0.06549	-12.150	<2e-16	0.45
red_Bupa.Chile	-0.25529	0.05095	-5.010	5.44e-07	0.77
red_Condes	-0.42922	0.12141	-3.535	0.000407	0.65
red_Empresas.Banmédica	0.05260	0.03045	1.728	0.084038	1.05

B.5. Resultados del Modelo de Regresión Múltiple

Tabla 2: Modelo Final RLM

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	0.0563	0.0028	19.782	$< 2 \times 10^{-16}$
deducibleUF	-0.8584	0.0052	-164.846	$< 2 \times 10^{-16}$
CopagoUF	0.5313	0.0044	119.619	$< 2 \times 10^{-16}$
ValorPrest	-0.1377	0.0045	-30.732	$< 2 \times 10^{-16}$
ValBonif	0.1243	0.0045	27.423	$< 2 \times 10^{-16}$
Año_2021	-0.0094	0.0013	-7.039	1.94×10^{-12}
Año_2022	-0.0116	0.0016	-7.354	1.94×10^{-13}
Año_2023	-0.0106	0.0016	-6.721	1.81×10^{-11}
Año_2024	-0.0101	0.0017	-6.112	9.86×10^{-10}
Mes_6	0.0019	0.0011	1.727	0.0842
Mes_10	0.0017	0.0010	1.740	0.0819
comunas1_LA.FLORIDA	0.0286	0.0010	27.551	$< 2 \times 10^{-16}$
comunas1_LAS.CONDES	0.0345	0.0012	28.143	$< 2 \times 10^{-16}$
comunas1_MAIPU	-0.0153	0.0029	-5.201	1.99×10^{-7}
comunas1_Otros	0.0079	0.0021	3.767	0.0002
comunas1_SANTIAGO	0.0063	0.0015	4.188	2.82×10^{-5}
comunas1_VITACURA	0.0259	0.0020	12.710	$< 2 \times 10^{-16}$
reg_XIII.REGION.METROPOLITANA	0.0088	0.0021	4.130	3.62×10^{-5}
rel_Otros	-0.0248	0.0115	-2.161	0.0307
rel_Titular	-0.0059	0.0006	-10.219	$< 2 \times 10^{-16}$
GENERO_M	0.0083	0.0006	14.598	$< 2 \times 10^{-16}$
I_Agrupado_Isapre	0.0165	0.0007	23.841	$< 2 \times 10^{-16}$
I_Agrupado_Otros	-0.0060	0.0019	-3.069	0.0021
red_Andes.Salud	-0.0110	0.0030	-3.619	0.0003
red_Bupa.Chile	-0.0126	0.0015	-8.523	$< 2 \times 10^{-16}$
red_Condes	0.0652	0.0019	34.925	$< 2 \times 10^{-16}$
red_Empresas.Banmédica	-0.0084	0.0008	-10.558	$< 2 \times 10^{-16}$
red_Grupo.Alemana	0.0700	0.0022	31.145	$< 2 \times 10^{-16}$
red_Meds	0.0142	0.0028	5.040	4.65×10^{-7}
red_Red.Interclínica	-0.0152	0.0025	-5.966	2.44×10^{-9}
red_UC	0.0385	0.0018	21.485	$< 2 \times 10^{-16}$
Servicio_Cardiologia	0.0102	0.0025	4.070	4.71×10^{-5}
Servicio_Dermatologia	0.0257	0.0016	16.276	$< 2 \times 10^{-16}$

B.6. Resultados del Modelo de Regresion Poisson

Tabla 3: Modelo Final RLP

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4940	0.0409	-12.068	$< 2 \times 10^{-16}$
Año2022	-0.4921	0.0107	-46.054	$< 2 \times 10^{-16}$
Año2023	-0.3817	0.0102	-37.300	$< 2 \times 10^{-16}$
Año2024	-0.2446	0.0115	-21.318	$< 2 \times 10^{-16}$
Mes2	-0.0670	0.0099	-6.804	1.02×10^{-11}
Mes3	-0.0332	0.0085	-3.914	9.06×10^{-5}
Mes4	-0.0360	0.0084	-4.262	2.02×10^{-5}
Mes5	-0.0743	0.0104	-7.117	1.10×10^{-12}
Mes6	-0.0731	0.0102	-7.157	8.27×10^{-13}
Mes7	-0.0675	0.0103	-6.567	5.15×10^{-11}
Mes9	-0.0347	0.0102	-3.405	0.0007
Mes11	0.0474	0.0094	5.051	4.39×10^{-7}
relHijo/a	0.0562	0.0084	6.732	1.68×10^{-11}
relOtros	-0.8405	0.1063	-7.908	2.63×10^{-15}
relTitular	0.4329	0.0076	57.177	$< 2 \times 10^{-16}$
I_AgrupadoIsapre	0.0605	0.0057	10.539	$< 2 \times 10^{-16}$
I_AgrupadoOtros	-0.9728	0.0177	-54.867	$< 2 \times 10^{-16}$
regXIII REGION METROPOLITANA	-0.2078	0.0185	-11.230	$< 2 \times 10^{-16}$
comunas1LA FLORIDA	0.7551	0.0226	33.410	$< 2 \times 10^{-16}$
comunas1LAS CONDES	0.0734	0.0235	3.120	0.0018
comunas1MAIPU	-0.1795	0.0338	-5.310	1.10×10^{-7}
comunas1MELIPILLA	-0.2311	0.0408	-5.665	1.47×10^{-8}
comunas1Otros	0.3275	0.0272	12.037	$< 2 \times 10^{-16}$
comunas1PROVIDENCIA	0.8022	0.0207	38.783	$< 2 \times 10^{-16}$
comunas1RECOLETA	0.2977	0.0248	11.982	$< 2 \times 10^{-16}$
comunas1SAN BERNARDO	-0.1033	0.0393	-2.631	0.0085
comunas1SANTIAGO	0.4919	0.0233	21.146	$< 2 \times 10^{-16}$
comunas1VITACURA	0.1032	0.0256	4.037	5.42×10^{-5}
ServicioCardiologia	-0.6323	0.0254	-24.878	$< 2 \times 10^{-16}$
ServicioDermatologia	-0.2649	0.0170	-15.544	$< 2 \times 10^{-16}$
ServicioDesconocidos	-0.2023	0.0166	-12.220	$< 2 \times 10^{-16}$
ServicioEndocrinología	-0.4144	0.0244	-16.964	$< 2 \times 10^{-16}$
ServicioGastroenterología	-0.4841	0.0234	-20.676	$< 2 \times 10^{-16}$
ServicioGeneral	0.6278	0.0121	51.719	$< 2 \times 10^{-16}$
ServicioGinecología	0.1447	0.0141	10.248	$< 2 \times 10^{-16}$
ServicioMedicina Familiar	-0.3603	0.0259	-13.905	$< 2 \times 10^{-16}$
ServicioMedicina Interna	-0.3303	0.0193	-17.112	$< 2 \times 10^{-16}$
ServicioNeurologia	-0.4316	0.0198	-21.813	$< 2 \times 10^{-16}$
ServicioOtros	-0.1207	0.0133	-9.106	$< 2 \times 10^{-16}$
ServicioOtorrinolaringología	-0.2705	0.0171	-15.843	$< 2 \times 10^{-16}$
ServicioPediatria	0.6028	0.0154	39.139	$< 2 \times 10^{-16}$
ServicioPsicologia	-0.3879	0.0282	-13.736	$< 2 \times 10^{-16}$
ServicioRespiratorias	-0.6925	0.0297	-23.319	$< 2 \times 10^{-16}$
ServicioReumatología	-0.7059	0.0332	-21.274	$< 2 \times 10^{-16}$
ServicioTraumatología	0.1758	0.0136	12.935	$< 2 \times 10^{-16}$
ServicioUrología	-0.3715	0.0200	-18.619	$< 2 \times 10^{-16}$
GENEROF	0.8646	0.0167	51.764	$< 2 \times 10^{-16}$
GENEROM	0.8591	0.0167	51.432	$< 2 \times 10^{-16}$
BonIsapreUF1	0.1420	0.0097	14.599	$< 2 \times 10^{-16}$
CopagoUF1	0.2322	0.0098	23.636	$< 2 \times 10^{-16}$
ValorPrestacionUF1	-0.1311	0.0096	-13.607	$< 2 \times 10^{-16}$

B.6.1. Predictoras Random Forest Reembolso

Tabla 4: Predictoras Random Forest Reembolso

Variables	Primera Capa	Mes	Red de Prestadores
Relación	Año	Comuna	Genero
Deducibleuf	Copagouf	Valorprest	Valbonif
Servicio			

B.6.2. Predictoras Boosted Trees Reembolso

Tabla 5: Predictoras Boosted Trees Reembolso

Variables	Primera Capa	Mes	Año
Comunas	Genero	Deducibleuf	Copagouf
Valorprest	Valbonif	Servicio	

B.6.3. Predictoras Random Forest Numero de Siniestros

Tabla 6: Predictoras Random Forest Numero de Siniestros

Variables	Primera Capa	Mes	Red de Prestadores
Relación	Año	Comuna	Genero
Servicio	Valorprest	Valbonif	

B.6.4. Predictoras Boosted Trees Numero de Siniestros

Tabla 7: Predictoras Boosted Trees Numero de Numero de Siniestros

Variables	Primera Capa	Mes	Año
Comunas	Genero	Valorprest	Copagouf
Valbonif	Región	Servicio	Relacion

Referencias

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211-243.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Cordeiro, M. F. M. (2023). A Machine Learning Approach to Predict Health Insurance Claims.

Dunn, P. K., Smyth, G. K., et al. (2018). *Generalized linear models with examples in R* (Vol. 53). Springer.

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman; Hall/CRC.

Frees, E. (2018). Loss data analytics. *arXiv preprint arXiv:1808.06718*.

Fundación MAPFRE. (2025). Seguro [Accedido el 3 de marzo de 2025]. https://www.fundacionmapfre.org/publicaciones/diccionario-mapfre-seguros/seguro/

García, J. F. S., Aguilar, D. S. G., Nieto, D. A. H., et al. (2023). Modelación de una prima de seguros mediante la aplicación de métodos actuariales, teoría de fallas y Black-Scholes en la salud en Colombia. *Revista de Metodos Cuantitativos para la Economia y la Empresa*, *35*, 330-359.

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric Statistical Inference* (5th). CRC Press.

Gómez Déniz, E., & Sarabia Alegría, J. M. (2008). *Teoría de la credibilidad: desarrollo y aplicaciones en primas de seguros y riesgos operacionales*. Fundación Mapfre Guanarteme.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Graziadei, H., de Melo, E. F., Targino, R. S., et al. (2023). Conformal prediction for frequency-severity modeling. *arXiv preprint arXiv:2307.13124*.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? https://arxiv.org/abs/2207.08815

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659-3667.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Ho, T. K. (1995). Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 278-282.

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss Models: From Data to Decisions*. John Wiley & Sons.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23.

López, A. (2004). Modelos lineales generalizados. *Universidad de Valencia, Burjassot, Valencia*.

Montgomery, D. C. (2019). *Design and Analysis of Experiments* (10th). John Wiley & Sons.

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th). John Wiley & Sons.

Peña, D. (2002). *Análisis de Datos Multivariantes* (1ra). McGraw-Hill.

Piontkowski, J. (2020). Forecasting health expenses using a functional data model. *Annals of Actuarial Science*, *14*(1), 72-82.

Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims. *Risks*, *11*(9), 164.

R Core Team. (2024). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd). Wiley.

Rincón, L. (2012). Introducción a la teoría del riesgo. *México: Facultad de Ciencias, UNAM*.

Shwartz-Ziv, R., & Armon, A. (2021). Tabular Data: Deep Learning is Not All You Need. https://arxiv.org/abs/2106.03253

(SII). (2024). *Nómina de Personas Jurídicas Inscritas en el Registro de Personas Jurídicas Sin Fines de Lucro* [Consultado el 29 de octubre de 2024]. https://www.sii.cl/sobre_el_sii/nominapersonasjuridicas.html

Terven, J. R. (2023). Loss Functions and Metrics in Deep Learning [arXiv preprint (version del 5 de julio de 2023)].

Tuininga, F. (2022). *A machine learning approach for modeling frequency and severity* [Tesis de maestría, University of Twente].

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3/4), 330-336. https://doi.org/10.1093/biomet/38.3-4.330

Werner, G., & Modlin, C. (2016). *Basic Ratemaking* (5th). Casualty Actuarial Society.

Zhang, Y., Dukic, V., & Guszcza, J. (2012). A Bayesian non-linear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *175*(2), 637-656.