Maestría en Ciencias Actuariales

# Predicción Del Costo De Las Reclamaciones Con Machine Learning

Yenny Marcela Casanova Rincón

Bogotá, D.C., 20 de Mayo de 2022

Escuela Colombiana de Ingeniería Julio Garavito
Vigilada Mineducación

# Predicción Del Costo De Las Reclamaciones Con Machine Learning

Tesis para optar al título de Magíster en Ciencias Actuariales

Catalina Lozano Murcia

Bogotá, D.C., 20 de Mayo de 2022

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
VIGILADA MINEDUCACIÓN

La tesis de maestría titulada "Predicción Del Costo De Las Reclamaciones Con Machine Learning", presentada por Yenny Marcela Casanova Rincón, cumple con los requisitos establecidos para optar al título de Magíster en Ciencias Actuariales

Director de la tesis
Catalina Lozano Murcia
Jurado
Yesid Esteban Clavijo Penagos
Jurado
Leonardo Vélez

Bogotá, D.C., 28 de Mayo de 2022

# Agradecimientos

En primer lugar, deseo expresar mi agradecimiento a mis padres por ser mi pilar fundamental y la motivación principal para culminar con éxito mis metas propuestas, por guiarme durante los periodos de duda y formarme como una buena persona.

Gracias a mi famila y amigos que me han prestado un gran apoyo moral y humano, necesarios en los momentos difíciles.

Agradezco a los todos docentes que, con su conocimiento y apoyo, motivaron a desarrollarme como profesional en la Escuela Colombiana de Ingeniería Julio Garavito, en especial a mi directora de tesis que, con su experiencia, me oriento en el desarrollo del trabajo de grado.

**Resumen**

El costo de las reclamaciones futuras constituye un estudio importante para determinar la tasa pura de cualquier cobertura que ofrezca una aseguradora, para ello se debe analizar el comportamiento de las reclamaciones y determinar una tendencia específica que se pueda aplicar en el costo del seguro, es por ello que surge la oportunidad de utilizar modelos predictivos presentes en el Machine Learning que ayuden a optimizar procesos y tener una noción más amplia de las reclamaciones. En esta ocasión se aplican diversos modelos de Machine Learning a una base de datos artificial cuyo comportamiento se apoya en siniestros reales asociados a accidentes laborales. Esta base es dividida en muestras de entrenamiento y validación, con el fin de entrenar los modelos sugeridos y validar la capacidad de predicción aplicando medidas a diferentes escenarios de manipulación de datos. Dentro de los resultados obtenidos se encuentra que el análisis descriptivo de datos da un buen indicio de la relación presente entre las características de las reclamaciones y el costo de la reclamación, el costo de las reclamaciones puede variar ampliamente dependiendo del modelo predictivo que se aplique y los niveles de ajuste varían de acuerdo al manejo que se realice en la base de datos.

# Índice

# Índice de cuadros

# Índice de figuras

# 1. Introducción

En el sector asegurador, una de las tareas más importantes del actuario es la tarifación de seguros, específicamente, cuantificando el riesgo asociado a un suceso impredecible como lo es la muerte, un accidente, enfermedad y demás causas que provoquen algún tipo de reclamo que conlleve a la indemnización del asegurado. El valor que resulta de esta medición, es la tasa pura o bruta de riesgo.
Uno de los factores necesarios para el cálculo de la tasa pura es el costo de las reclamaciones, la cual puede estimarse a futuro teniendo en cuenta la experiencia propia de la aseguradora, recopilando información de siniestralidad y realizando los cálculos pertinentes. Bajo este esquema, usualmente se utilizan modelos lineales generalizados (GLM) para la proyección del costo de las reclamaciones, dependiendo de parámetros como la edad, género y demás que permitan una apropiada proyección de la misma, sin embargo, en la actualidad existe el auge por el manejo de las grandes bases de datos y por tanto se han desarrollado nuevas técnicas computacionales basados en modelos de regresión que no están limitados a encontrar una relación lineal entre las variables dependientes e independientes entre los cuales encontramos los modelos de ***Machine Learning***, permitiendo la agilidad de procesos y una automatización de los mismos; y es por esto que resulta la oportunidad de innovar a la hora de predecir costos de reclamaciones.

La maestría en ciencias actuariales de la Escuela Colombiana de Ingeniería Julio Garavito tiene por objetivo dotar al estudiante de conceptos actuariales y a su vez, del manejo y análisis de datos, haciendo notar el complemento de la última en las funciones de un actuario. El objetivo general del trabajo es aplicar estos dos conceptos teóricos vistos en la maestría con el fin de resolver un problema de la vida real, concretamente, la proyección de costo de siniestros, además de aplicar nuevas metodologías de predicción en el sector asegurador.

Inicialmente se explica el funcionamiento de los modelos Machine Learning aplicados a la base de datos para predecir el costo de las reclamaciones, seguido de detallar las medidas utilizadas para la medición y comparación de los modelos resultantes, posteriormente se describe la base de datos a utilizar, analizando el comportamiento de cada variable y la relación existente entre las variables independiente y la variable a predecir, es decir, el costo de las reclamaciones. Finalmente se divide la base de datos en un conjunto de entrenamiento y otro de validación, necesarios para entrenar cada uno de los modelos sobre 4 escenarios propuestos resultantes de manipular la base de siniestros, estos modelos ya entrenados se ejecutan sobre las bases de validación, comparando los niveles de ajuste de cada uno por medio del error absoluto medio y el R-cuadrado.

# 2. Marco Teórico

En los últimos años, el aumento de información almacenada en grandes bases de datos ha hecho que la innovación de técnicas en análisis de datos haya evolucionado hasta el punto de crear modelos de aprendizaje autónomos que ayuden a la regresión y clasificación de datos. Estas técnicas avanzadas de computación se denominan Machine Learning, contenidas en el mundo de la Inteligencia Artificial, y han tomado cada año más relevancia en varios sectores económicos como salud, automóviles, ventas, publicidad, bancos, servicios financiero y aseguradoras como se muestra en [Fortune. (2022)], ayudando a la toma de decisiones, optimización de procesos, ahorro de costos entre otros beneficios.
El sector asegurador cuenta con diversas áreas que aprovechan las técnicas del Machine Learning, según lo detalla el artículo [Insurance. (2021)] que hace referencia a la encuesta realizada por la compañía SMA especificando que en 2019, las aseguradoras determinaron que el Machine Learning es altamente relevante dentro de las funciones actuariales y análisis del riesgo de suscripción. Dentro de las funciones actuariales, el estudio [Blier-Wong, C. (2020) ] especifica que la técnicas de Machine Learning han sido utilizadas en mayor medida en Pricing o tarificación de seguros, es por ello que se plantea utilizar los modelos de regresión supervisada a un tema específico de Pricing como lo es la estimación de costos por reclamaciones, este caso, asociados a la cobertura de accidentes laborales. Estos modelos tienen por finalidad "aprender" de un conjunto de datos previamente etiquetados permitiendo que los resultados arrojados por el modelo sean comparables con los datos conocidos de la base. Los modelos propuestos para lograr dicha predicción son descritos brevemente tomando como referencia la teoría detallada presente en [Miller, D. (2017)] y [Ramasubramanian, K. (2019)], así como los métodos para determinar el nivel de error y predicción de los modelos resultantes.

## 2.1. Modelos de regresión supervisada

### 2.1.1. Regresión Lineal Múltiple

La regresión lineal múltiple permite encontrar un modelo cuya relación entre la variable dependiente ($Y$) y el conjunto de variables independientes o predictoras ($X_1, X_2, X_3 \dots$) sea lineal, es decir, la relación entre la variable dependiente y el conjunto de variables independientes se rige la ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + \epsilon_i \tag{1}$$

Donde:

- $\beta_0$: El valor de la variable dependiente $Y$, cuando todos los predictores son cero.
- $\beta_i$: Es el efecto promedio que tiene la variable predictora $X_i$ sobre la variable dependiente $Y$ Se conocen como coeficientes parciales de regresión.

- $\epsilon_i$: Es el residuo o error. La diferencia entre el valor observado y el estimado por el modelo.

Para que este modelo sea insesgado (los resultados sean parecidos a los reales) óptimo (varianza mínima) es necesario cumplir con los siguientes requisitos:

- **No multicolinealidad**: Independencia lineal entre las variables predictoras. La multicolinealidad sucede cuando un predictor es una combinación lineal de una o varias variables predictoras.

- **Relación lineal entre la variable dependiente y los predictores:** Es claro que para generar un modelo lineal, la relación entre la variable a predecir y las variables predictoras debe ser lineal, una forma aproximada de validar esta condición es analizando la gráfica de los residuos del modelo versus cada uno de los valores ajustados (**Figura 1**) . Si los residuos se distribuyen de forma aleatoria en torno a cero (**Figura 1, Caso 1**), se puede dercir que existe una relación lineal entre la variable independiente y los predictores. Si, por el contrario, los residuos empiezan a repartirse generado una forma específica (no aleatoria) (**Figura 1, Caso 2**) nos encontramos con el que modelo no es el mejor, ya que la relación entre la variable dependiente y las independientes no es lineal; podría ser cúbica, logarítmica, exponencial, entre otras.

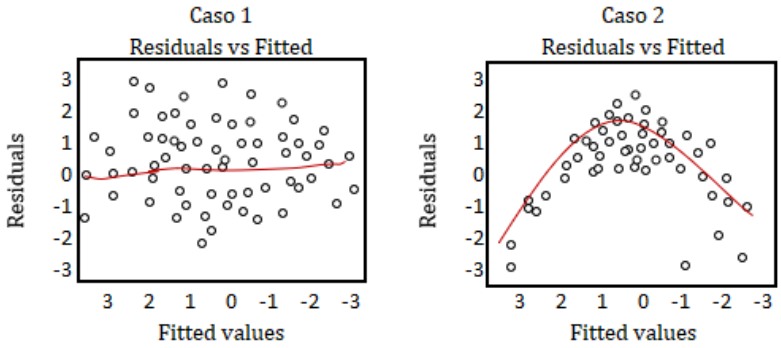


Figura 1: Residuos Vs Valores Ajustados

- **Distribución normal de los residuos:** Corroborar que los residuos se distribuyen de forma normal garantiza que la variable predicha, $Y$, dependa linealmente de los coeficientes de regresión $\beta_i$. Para validar que los residuos se distribuyen de forma normal, se hace uso del gráfico **Cuantil-Cuantil** que, en este caso, se compara los cuantiles teóricos de una distribución normal, vs los cuantiles de los residuos normalizados (**Figura 2**). Si se nota una tendencia lineal (**Figura 2, Caso 1**), los residuos distribuyen similar a la distribución normal , de lo contrario (**Figura 2, Caso 2**), podríamos decir que no sigue una distribución normal.

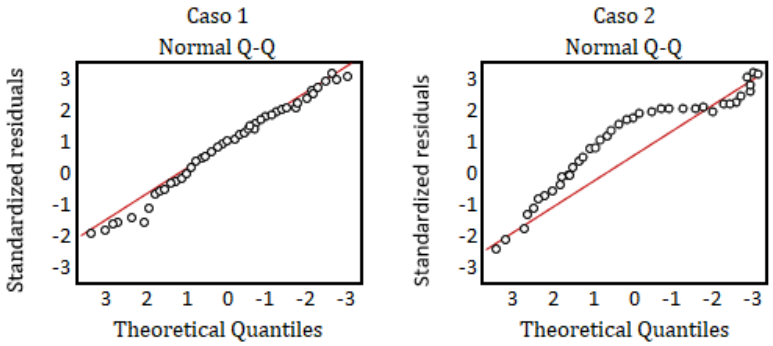


Figura 2: Gráficos Cuantil-Cuantil

- **Homocedasticidad:** Otros de los criterios importantes para determinar la validez del modelo lineal es la homocedasticidad, es decir, que la varianza de los residuos sea constante en todo rango de observación. El método gráfico para validar la homocedasticidad es el gráfico de "Escala-Ubicación" (**Figura 3**) que consiste en plasmar el comportamiento de la raíz cuadrada de los residuos normalizados frente los valores ajustados. Si los datos representados parece no seguir alguna tendencia o distribución (**Figura 3, Caso 1**), se puede concluir que existe homocedasticidad en los residuos, de lo contrario (**Figura 3, Caso 2**), la varianza de los residuos no es constante para todo rango de observación.

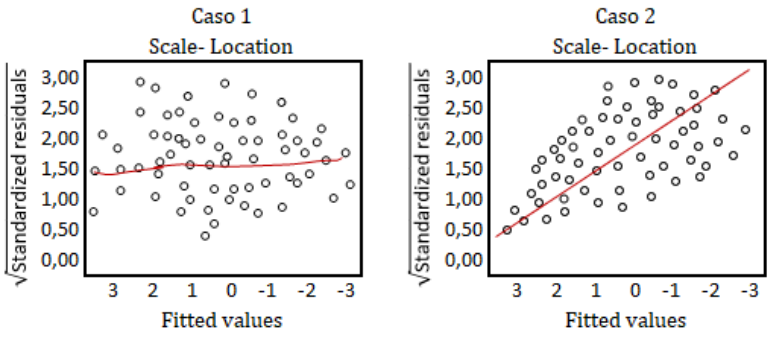


Figura 3: Gráficos: Escala-Ubicación

- **Valores atípicos:** Es importante identificar observaciones atípicas que puedan sesgar el modelo, para ello se emplea el gráfico de residuales vs apalancamiento con la distancia de Cook. Dentro de la gráfica podemos observar unas líneas punteadas que representa la distancia de Cook, los puntos dentro de esta área podemos decir que son puntos que no sesgan el modelo (**Figura 4, Caso 1**), mientras que los puntos fuera de esta área punteada, al extremo interior o superior derecho

(**Figura 4, Caso 2**), pueden catalogarse como puntos atípicos u outliers que sesgan el modelo lineal.

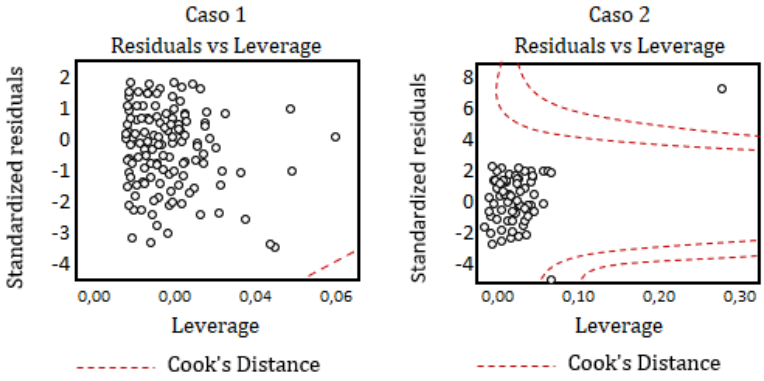


Figura 4: residuales vs palancamiento (leverage)

Fuera de los requisitos necesarios para determinar la validez del modelo lineal, también se puede identificar la influencia que tienen los predictores sobre el modelo lineal resultante, con denominado **P-valor** que comprueba la hipótesis nula de que el coeficiente que acompaña la variable predictora es igual a cero, es decir,

$$H_{0,i} : \beta_i = 0, \quad \text{para todo } i = 1, 2, \ldots, n \tag{2}$$

Por tanto, lo que se desea buscar para cada uno de los predictores, es un P-Valor bajo ($<0.05$) y rechazar la hipótesis nula.

### 2.1.2. Árboles de regresión

Los árboles de decisión para regresión guardan similitud con los diagramas de flujo, estableciendo un conjunto de reglas sucesivas que ayudan a tomar una decisión.

Dentro de la estructura de árbol, podemos encontrar los **Nodos Prueba** los cuales se particionan en nuevos nodos con una condición dada por el modelo, el primer nodo prueba del árbol tambien se denomina **Nodo Raíz**. Los nodos que no pueden ser particionados se denominan **Nodos de Decisión** o **Nodos Rerminales**.

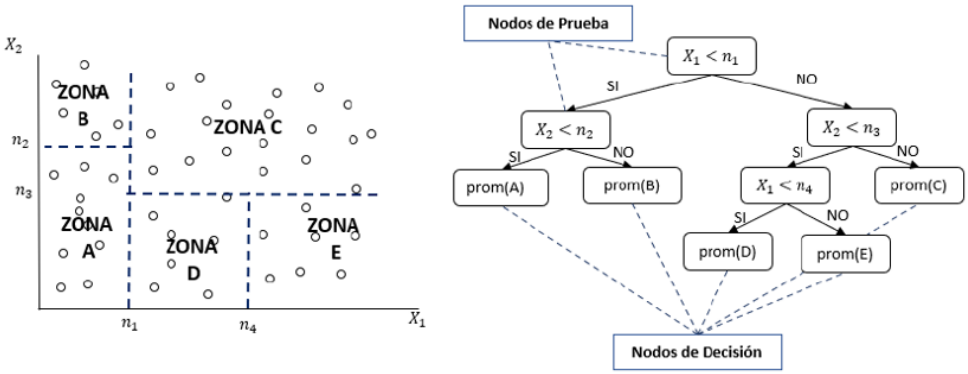


Figura 5: Árbol de decisión en $R^2$

Como se ve en la **Figura 5**, un árbol de decisión particiona el espacio formado por el conjunto de variables predictoras $X_i$ de acuerdo con una serie de condiciones que comparan las variables independientes respecto a un valor dado $n_j$. Para cada una de las particiones encontradas, el modelo promedia los valores pertenecientes a cada una de las zonas, permitiendo que las observaciones que lleguen al modelo tomen el valor $\hat{y}_i$ que corresponde al promedio de la zona en la cual pertenece la observación.

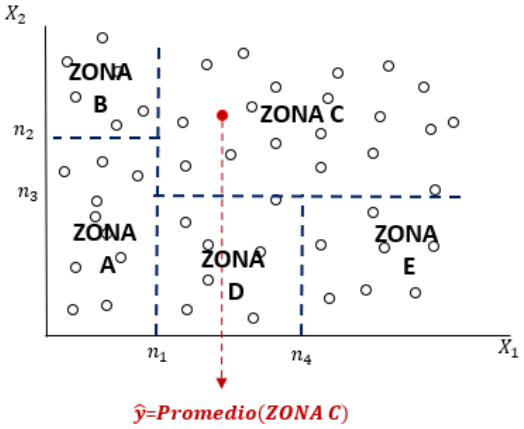


Figura 6: Valor predicho por el árbol de decisión

Este proceso se repite consecutivamente hasta encontrar la partición que permita minimizar la función de costos:

$$\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2 \tag{3}$$

### 2.1.3. Redes Neuronales

Las Redes neuronales o sistemas conexionistas son un modelo algorítmico que se basa en un conjunto de unidades neuronales simples llamadas neuronas artificiales, haciendo que su funcionamiento sea similar a los axones de las neuronas que tiene cualquier ser organismo biológico.

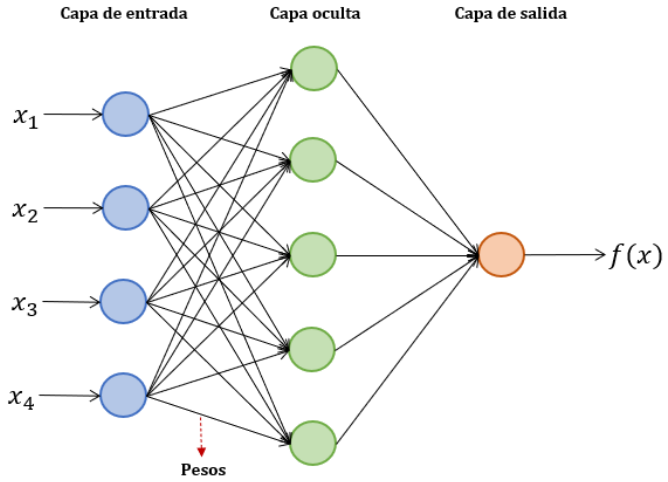


Figura 7: Esquema de la Red Neuronal

Dentro del esquema de la red neuronal podemos encontrar la capa de entrada, encargada de recibir las variables predictoras $x_i$ en bruto (sin ningún proceso que modifique su valor), la capa oculta (o capas ocultas) pondera los valores que le entrega la capa de entrada de acuerdo a las conexiones (flechas) de cada variable predictora, finalmente la capa de salida combina los valores que resultan de la capa oculta para generar las predicciones $\hat{y}$.

Estos sistemas van aprendiendo y se forman a sí mismos en lugar de ser programados de forma explícita, por tanto, se les suele llamar las cajas de negras de la clasificación.

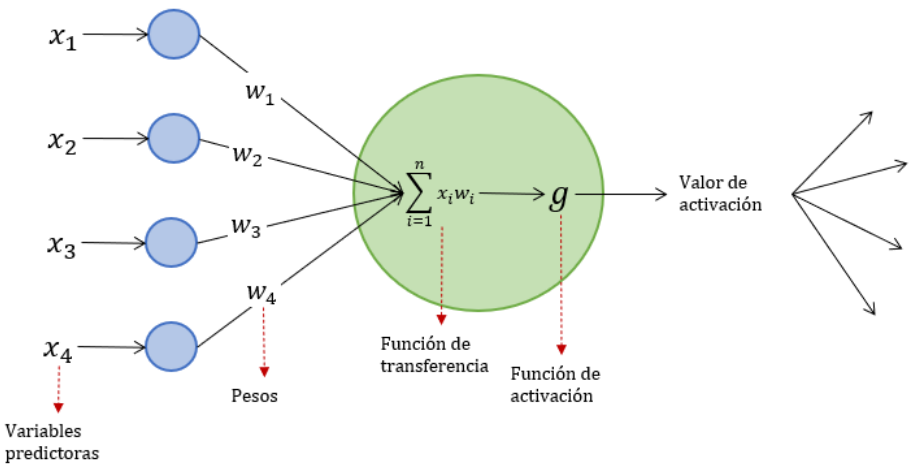


Figura 8: Detalle de la Red Neuronal

Haciendo "Zoom" en el procedimiento de cada red neurona (**Figura 8**), una vez llegan los valores de la capa de entrada a la neurona, esta pondera los valores obtenidos por medio de la **Función de Transferencia** de acuerdo a los pesos, $w_i$, que determinarán la fuerza o importancia de cada conexión, el valor resultante es transmitido a la **Función de activación** la cual controla que información se propaga desde una capa hacia otra, cuando esta función toma el valor de 0 se dice que la neurona está inactiva, pues no transfiere ningún valor. Una vez superado el umbral de la función de activación, el valor que se propaga a las demás neuronas, se denomina **Valor de Activación**.

### 2.1.4. Bagging

También llamado **Bootstrap Aggregation** consiste en entrenar los diferentes modelos con subconjuntos del conjunto de entrenamiento, dando a cada resultado obtenido un peso para su ponderación. Por defecto, las muestras escogidas para cada aprendiz son escogidas con reemplazo, por tanto, una muestra dada puede ser escogida más de una vez en cada bloque de entrenamiento aleatorio. Los modelos obtenidos con cada subconjunto, se ensamblan creando un único modelo predictivo, que es el promedio de todos los modelos.

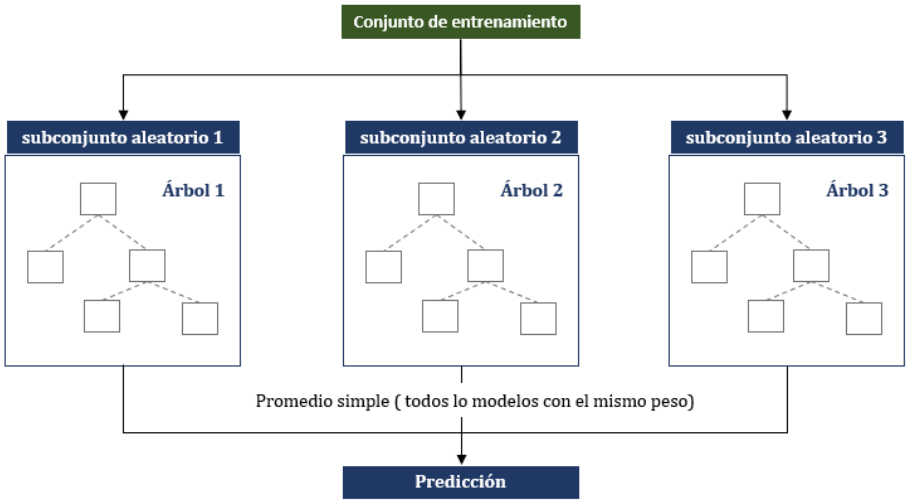


Figura 9: Esquema del modelo Bagging

Estos modelos se aplican por lo generar a modelos con alta variabilidad en sus resultados, ya que su finalidad es disminuir la varianza del modelo con el remuestreo sucesivo aleatorio. En este caso, se hizo uno del modelo de Bagging entrenando arboles de decisión.

### 2.1.5. Gradient Boosting

Este método de ensamble o de combinación de modelos permite aprender del error del pasado, es decir, Boosting es una forma técnica de aprendizaje secuencial, cada modelo resultante emplea información del modelo anterior con la finalidad de aprender de los errores y disminuir gradualmente el margen de error respecto a la iteración anterior. En este caso se utiliza el modelo de árboles de decisión para su mejora a través del modelo Boosting, que a diferencia del Bagging, este no hace un muestreo repetitivo, sino que toma todo el conjunto de entrenamiento dándole en la siguiente iteración un peso mayor en las variables cuya predicción fue la menos acertada. El resultado final será la ponderación de los resultados de cada árbol de regresión cuyos pesos son las puntuaciones de precisión de los resultados de cada regresión.

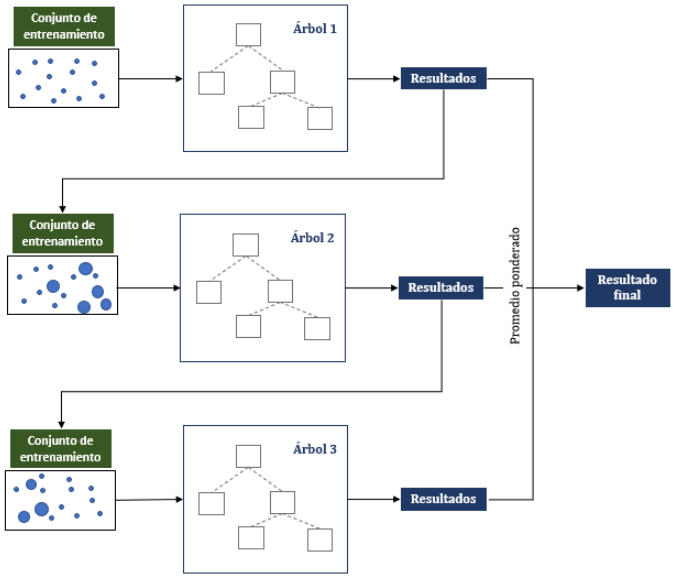


Figura 10: Esquema del modelo Boosting

Para determinar el nivel de error y ajuste de los modelos resultantes, se propone las siguientes técnicas de medición.

## 2.2. Metodos de medición

### 2.2.1. Error Absoluto Medio (EAM)

La finalidad es cuantifica el error de los modelos de predicción comparando los valores reales frene los valores predichos. Esta medida se define como sigue:

$$EAM = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{4}$$

donde:

$y_i$ son los valores que toma la variable dependiente real.

$\hat{y}_i$ son los valores predichos por el modelo.

$n$ es el número de observaciones de la predicción.

### 2.2.2. R-Cuadrado

El R-cuadrado o coeficiente de determinación mide la bondad de ajuste del modelo a la variable que quiere predecir o explicar. Se define como sigue:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

donde:

$y_i$ son los valores que toma la variable dependiente real.

$\hat{y}_i$ son los valores predichos por el modelo.

$\bar{y}$ es la media muestral de la variable independiente real $Y$.

$n$ es el número de observaciones de la predicción.

# 3. Metodología

Considerando que la idea principal es aplicar modelos supervisados de Machine Learning para estimar el costo de las reclamaciones asociados a alguna cobertura existente en el mercado asegurador. Inicialmente se toma una base de datos que represente las reclamaciones asociadas a una cobertura específica. Una vez obtenida la base de datos, esta es cargada y manipulada en la herramienta de programación R, donde se realiza un análisis preliminar del comportamiento de cada una de las variables, inspeccionando valores faltantes y atípicos, y encontrado posibles relaciones existentes entre las variables independientes y la variable a predecir, es decir, el costo de la reclamación. Se proponen 4 escenario que modifican la base de datos original, para cada uno de estos escenarios se toma una muestra de entrenamiento o test correspondiente al 70 % de los registros, que permitirá ajustar cada uno de los modelos propuestos previamente al escenario. Una vez obtenido el modelo, se usa la muestra de validación la cual es el 30 % restante de la base (diferentes a los datos de entrenamiento) para aplicar el modelo y contrastar los valores resultantes con los valores reales presenten en la muestra de validación, para logras este comparativo se utiliza el error absoluto definido (EAM) y el R-cuadrado.

## 3.1. Estructura de la base de datos

La base para el estudio fue tomada del repositorio de Kaggle que incluye 36146 pólizas de seguros de accidentes laborales. La base de datos es sintética, sin embargo, los registros son semejantes al comportamiento de las reclamaciones por accidentes laborales presentadas a una aseguradora donde cada registro presenta información demográfica y relacionada con el trabajador, así como una descripción de texto del accidente. La información se detalla a continuación:

Cuadro 1: Descripción de la tabla

| VARIABLE | DESCRIPCIÓN | CLASE | FORMATO |
|---|---|---|---|
| **ClaimNumber** | Identificación del reclamo | chr | "WC8205482"... |
| **DateTimeOfAccident** | Fecha en la que ocurrió el siniestro | POSIXct | 2002-04-09T07:00:00Z |

Cuadro 1: Descripción de la tabla

| VARIABLE | DESCRIPCIÓN | CLASE | FORMATO |
|---|---|---|---|
| **DateReported** | Fecha del reporte del siniestro | POSIXct | 2002-07-05T00:00:00Z |
| **Age** | Edad del asegurado en el momento del siniestro | num | 48, 43,30, 41, 36... |
| **Gender** | Genero del asegurado | chr | M: Masculino<br>F: Femenino |
| **MaritalStatus** | Estado civil de quien reclamó | chr | M: Casado<br>U: Unión libre<br>S: Soltero |
| **DependentChildren** | Número de hijos a cargo | num | 0,1,2,3,4... |
| **DependentsOther** | Otras personas a cargo | num | 0,1,2,3,4... |
| **WeeklyWages** | Salario semanal | num | 500, 509, 709, 555, 377 ... |
| **PartTimeFullTime** | Si trabaja a tiempo completo | chr | F: Completo<br>P: Medio Tiempo |
| **HoursWorked-PerWeek** | Horas trabajadas a la semana | num | 38, 37.5, 38, 38, 38 ... |
| **DaysWorked-PerWeek** | Días trabajados por semana | int | 1, 2, 3, 4, 5... |
| **ClaimDescription** | Descripción del reclamo | chr | LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY |
| **InitialIncurred-CalimsCost** | Valor inicial del reclamo | num | 1500, 5500, 1700, 15000... |
| **UltimateIncurred-ClaimCost** | Costo Total del reclamo | num | 4748, 6326, 2294, 17786... |

## 3.2. Preprocesamiento de datos

Como es habitual, antes de implementar un modelo de predicción, es necesario conocer la base de datos en cuanto su estructura, interpretabilidad, valores máximos, mínimos, frecuencias y demás factores ayuden a sacar conclusiones sobre la información. Empezamos generando un resumen por cada una de las variables y analizando inicialmente cuáles variables pueden ser descartadas de estudio por su poca relevancia con el costo del siniestro.

Cuadro 2: Resumen inicial de la base de datos

| DateTimeOfAccident | DateReported | Age |
|---|---|---|
| Min. :1988-01-01 09:00:00 | Min. :1988-01-10 00:00:00 | Min. :13.0 |
| 1st Qu.:1992-07-03 10:00:00 | 1st Qu.:1992-08-10 00:00:00 | 1st Qu.:23.0 |
| Median :1996-12-26 05:00:00 | Median :1997-02-09 00:00:00 | Median :32.0 |
| Mean :1996-12-30 10:12:51 | Mean :1997-02-07 01:20:23 | Mean :33.8 |
| 3rd Qu.:2001-07-05 12:00:00 | 3rd Qu.:2001-08-21 00:00:00 | 3rd Qu.:43.0 |
| Max. :2005-12-31 10:00:00 | Max. :2006-09-23 00:00:00 | Max. :79.0 |

| DaysWorkedPerWeek | InitialIncurredCalimsCos | UltimateIncurred-ClaimCost |
|---|---|---|
| 1 : 119 | Min. : 1 | Min. : 122 |
| 2 : 336 | 1st Qu.: 700 | 1st Qu.: 925 |
| 3 : 956 | Median : 2000 | Median : 3373 |
| 4 : 995 | Mean : 7744 | Mean : 10951 |
| 5: 32969 | 3rd Qu.: 9500 | 3rd Qu.: 8185 |
| 6: 563 | Max. :830000 | Max. :4027136 |
| 7: 208 | | |

| HoursWorked-PerWeek | WeeklyWages | DependentChildren | DependentsOther |
|---|---|---|---|
| Min. : 0.00 | Min. : 1.0 | 0 : 33891 | 0 : 35851 |
| 1st Qu.: 38.00 | 1st Qu.: 200.0 | 1 : 858 | 1 : 297 |
| Median : 38.00 | Median : 393.3 | 2 : 924 | 2 : 15 |
| Mean : 37.77 | Mean : 416.4 | 3 : 353 | 3 : 6 |
| 3rd Qu.: 40.00 | 3rd Qu.: 500.0 | 4 : 103 | |
| Max. :640.00 | Max. :7497.0 | 5: 34 | |
| NA's :49 | NA's :56 | 6: 4 | |
| | | 8:1 | |
| | | 9:1 | |

| Variable Categórica | Categorías |
|---|---|
| ClaimNumber | 29438 |
| Gender | 2 |
| MaritalStatus | 3 |
| PartTimeFullTime | 2 |
| ClaimDescription | 20581 |

**Frecuencia por categoría**

| Gender | MaritalStatus | PartTimeFullTime |
|---|---|---|
| F:8250 | M:15159 | F:32887 |
| M:27896 | S:17448 | P:3259 |
| | U:3539 | |

De los datos obtenidos se realizan los siguientes ajustes sobre la base:

- La identificación de los siniestros, "ClaimNumber", y la descripción del reclamo ,"ClaimDescription", tienen un alto número categorías, considerando que la base tiene 36.146 observaciones, por tanto, estas variables serán descartadas del estudio pues la información que contienen no influyen en el costo del siniestro.

- La variable que indica el costo inicial del reclamo, en teoría, no aporta mayor información a la hora de predecir el costo total de un siniestro futuro.

- En lugar de trabajar directamente con la fecha de aviso, "DateReported', se crea una nueva variable denominada "days_Report" en la que se evidencia cuanto tiempo en días transcurrieron entre la ocurrencia del siniestro y el aviso del mismo a la aseguradora..

- La fecha de ocurrencia del siniestro,"DateTimeOfAccident" se va a transformar en dos nuevas variables "A.OCURRENCIA" y "M.OCURRENCIA". La primera representa el año de ocurrencia y la segunda el mes de ocurrencia del siniestro. Esto ayuda a evidenciar si existe alguna tendencia en el costo de las reclamaciones.

- Las variables "WeeklyWages" y "HoursWorkedPerWeek' presentan datos vacíos o NA's.

- Se corrobora que no existen vacíos ocultos en los datos tipo caracter, como espacios o categorías no relacionadas a la variable.

### 3.2.1. Imputación de datos

De los resumenes anteriores, las variables "WeeklyWages" y "HoursWorkedPerWeek' presentan datos vacíos o NA's los cuales se distribuyen así:

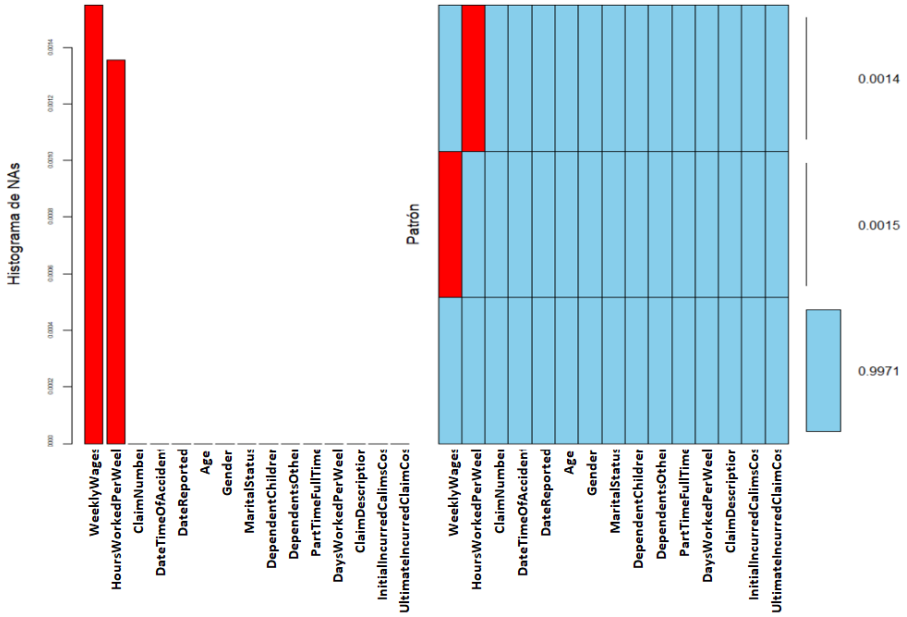


Figura 11: distribución de valores faltantes

La **Figura 11** nos muestra que las variables "WeeklyWages" y "HoursWorkedPer-Week" contienen datos vacíos, pero el conjunto de vacíos no se cruzan entre sí (una fila de observaciones no tiene más de un NA) y además el conjunto de datos faltantes no pesa más del 0.3 % del total de la base, lo cual, a la hora de hacer la imputación de datos no corremos un mayor riesgo de sesgar significativamente el modelo. por este motivo que se eliminan estos registros de la base de datos. Con esto se detalla que los datos están completos.

### 3.2.2. Análisis de datos

Una vez completa la base, se dispone a analizar el comportamiento de cada variable y la relación de que existe entre el costo del siniestro y las demás variables que serán las variables predictoras.
Inicialmente se detalla un punto atípico en la base de reclamos, lo mejor es prescindir de esta observación, pues podría sesgar de gran manera los modelos de predicción. una vez realizado este ajuste, se detalla en la **Figura 12** que aun prevalece un gran número de valores atípicos en la base de datos, donde el valor reclamado medio durante el periodo de observación es de 10845, y varía desde 122 hasta, 865770 unidades monetarias.

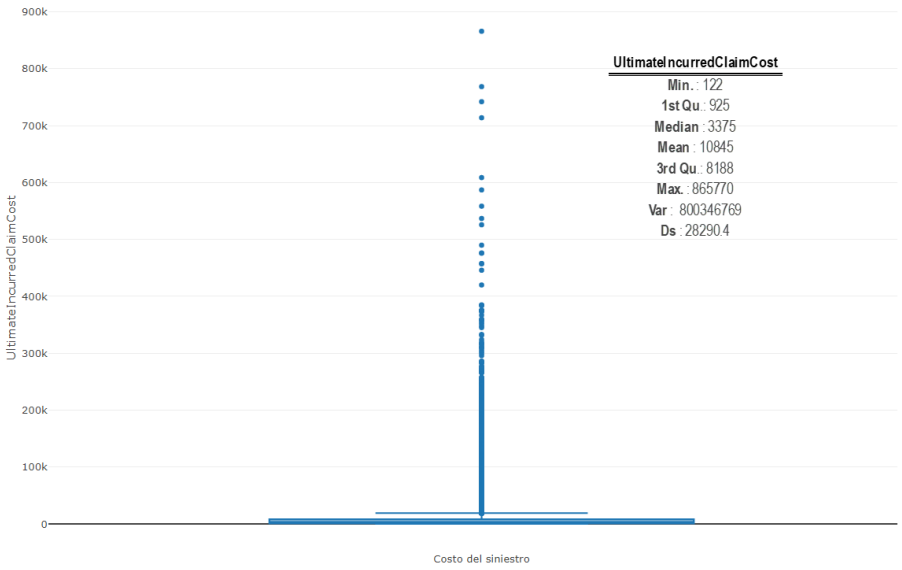


Figura 12: Boxplot del costo de las reclamaciones

Tambien es importante determinar que realción existe entre el costo de las reclamaciones y las caaracteristicas asosciadas al trabajador.

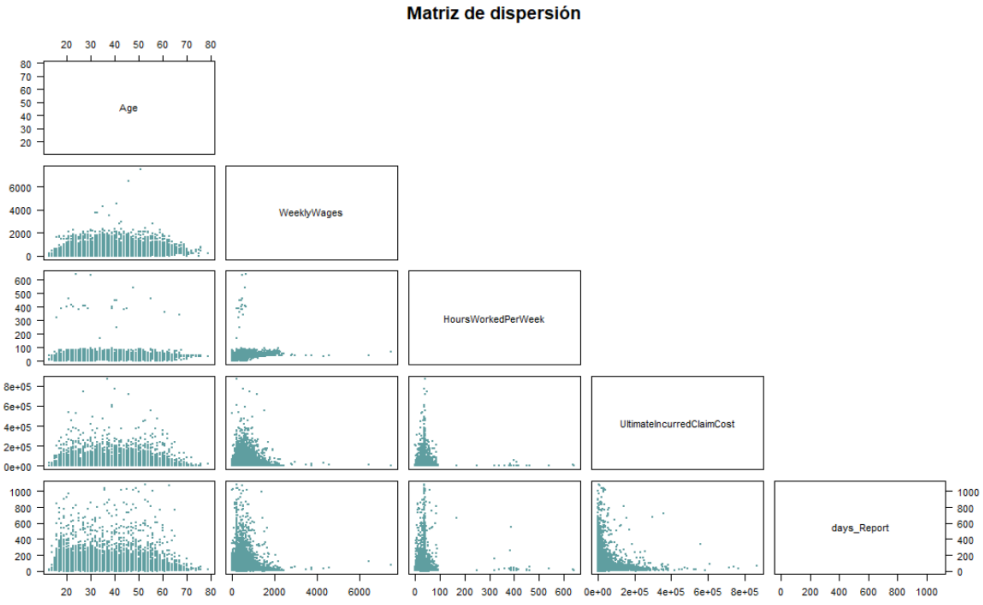


Figura 13: Matriz de dispersión entre variables numéricas

De la **Figura 13** se pueden conlcuir:

- Las variables no muestran un comportamiento lineal entre sí, esto se corrobora con la matriz de correlación.
- Los reclamos de mayor valor tienden a ser avisados más rápido que los de menor valor.
- Notamos que existen valores atípicos conceptualmente en las horas trabajadas por semana, ya que suponiendo que una persona trabaje 24 horas los 7 días a la semana, este lograría completar 168 horas de trabajo a la semana, y encontramos valores por encima de 200 horas de trabajo a la semana, lo cual implica pensar que existió un error operativo en la base. Los 27 registros que muestren horas semanales trabajadas por encima de 112 (correspondientes a 16 horas diarias por 7 días) serán eliminados de la base.
- Las personas de mayor edad tienden a trabajar menos días a la semana, lo cual conlleva a trabajar menos horas a la semana y, por tanto, a generar menores ingresos semanales. Así mismo el costo de la reclamación es menor.
- La relación entre el salario semanal y el costo del reclamo parece una distribución normal donde la mayor cantidad de observaciones está sesgada a menores salarios. Esta misma relación se puede ver entre las variables de los salarios semanales y los días de reporte del siniestro.
- Las variables numéricas, excepto la edad, presentan valores atípicos.

A continuación, se presenta el gráfico de costos medios respecto a cada una de las variables categóricas y tipo factor que existen:

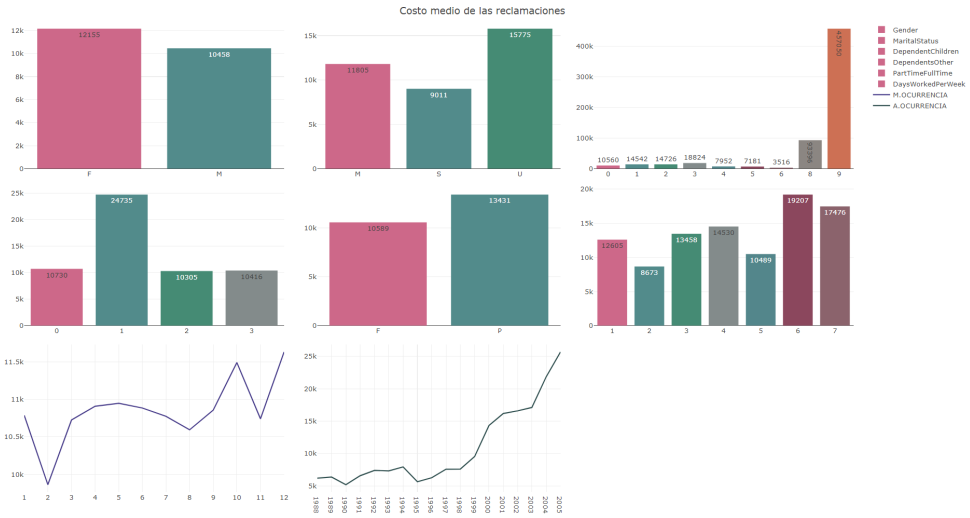


Figura 14: Costo medio de las reclamaciones

De las gráficas anteriores podemos notar lo siguiente:

- El costo medio de reclamación es superior en mujeres que en hombres, a pesar de que el porcentaje de participación de las mujeres en las reclamaciones es menor, esto quiere decir, que la siniestralidad en las mujeres es de baja frecuencia[1], pero alta severidad [2].

- El costo medio es mayor para las personas que están en unión libre, seguido de las personas casadas y por último las personas solteras. Las personas de unión libre presentan baja frecuencia, según el **cuadro 2**, y alta severidad.

- El costo medio de las reclamaciones hechas por las personas que tienen 9 hijos a cargo, es notoriamente más alto que la reclamación presentada por persona que tienen 5 hijos o menos. Esto parece un dato erróneo, pues según las tablas resumen anteriores, ese costo medio corresponde a un único reclamo, igualmente el costo medio asociado a la persona que tiene 8 hijos, así que estas dos observaciones se descartan, adicionalmente el caso de que una familia tenga 9 u 8 hijos es muy poco probable dadas las composiciones familiares actuales en la sociedad.

- El costo medio es mayor para las personas que tienen a cargo 1 persona diferente a un hijo.

- Quienes trabajan tiempo medio, tienden a presentar un reclamo con un valor mayor a los que trabajan tiempo completo.

- El costo medio de las personas que trabajan 6 días a la semana es más alto que los demás. Se resalta, además, que lo más común es que reclamen personas que trabajen 5 días a la semana.

- El costo medio tiende ser más alto en los meses de octubre y diciembre.

- Al pasar de los años, el costo medio de las reclamaciones ha aumentado significativamente.

Se realizan los ajustes en la base de datos según lo observado anteriormente, es decir, eliminar las observaciones que en las horas trabajadas por semana sean mayores de 112 horas, bajo el supuesto que lo máximo trabajado sean 16 horas por día, durante 7 días a la semana y las personas que tengan a cargo 8 o 9 hijos. En total se eliminaron 29 registros que corresponden al 0.08 % de las observaciones de la base original.

Con estos ajustes, la base muestra que a pesar de extraer las inconsistencias en la variable "HoursWorkedPerWeek", las variables siguen presentando datos atípicos. Adicionalmente el comportamiento de las variables categóricas se mantiene respecto a los costos medios, pero se detalla que quienes tienen 3 hijos a cargo, presentan reclamos a la aseguradora con un mayor costo que las demás personas.

[1] La frecuencia hace referencia al número de veces que se repite una observación durante la realización de un muestreo, en este caso, número de reclamaciones presentadas por mujeres sobre el total de reclamaciones.

[2] La severidad hace referencia al costo de la reclamaciones, en este caso quiere decir que el costo de las reclamaciones realizadas por las mujeres es más alta que la de los hombres.

## 3.3. Entrenamiento de modelos predictivos

Para determinar cuál de los modelos predictivos se ajusta mejor a los datos, se crean dos subbases para entrenar los modelos de regresión. El primero corresponde al 70 % de las observaciones manteniendo las características de la base original, esta base se denomina **conjunto de entrenamiento**, el 30 % restante será el **conjunto de validación**, donde se pondrá a prueba el modelo obtenido con le conjunto de entrenamiento con el fin de obtener el error de predicción del modelo resultante. Adicionalmente, se han creado 4 escenarios para el entrenamiento de cada uno de los modelos y así comparar que modelo se ajusta mejor y bajo qué condiciones, estos escenarios son:

- **Escenario 1:** Considerando la base de datos con los ajustes detallados hasta el momento y adicionando variables dummy para las variables categóricas. Las variables que conforman la base inicial son las siguientes:

Cuadro 3: Descripción base de datos del escenario 1

| variable | Descripción |
|---|---|
| **Age** | Edad de quien reclamó |
| **Gender.M** | Variable Dummy<br>1: Si el genero el masculino<br>0: En otro Caso |
| **MaritalStatus.M** | Variable Dummy<br>1: Si es casado<br>0: En otro Caso |
| **MaritalStatus.S** | Variable Dummy<br>1: Si es soltero<br>0: En otro Caso |
| **DependentChildren** | Número de hijos a cargo. |
| **DependentsOther** | Otras personas a cargo. |
| **WeeklyWages** | Salario semanal. |
| **PartTimeFullTime.F** | Variable Dummy<br>1: Si trabaja tiempo completo<br>0: En otro Caso |
| **HoursWorkedPerWeek** | Horas trabajadas a la semana |
| **DaysWorkedPerWeek** | Días trabajados por semana |
| **A.OCURRENCIA** | Año de ocurerncia del siniestro |
| **M.OCURRENCIA** | Mes de ocurrencia del siniestro |
| **days_Report** | Días transcurridos desde la fecha de ocurrencia y la fecha de aviso |

- **Escenario 2:** Tomando la base del escenario 1 excluyendo las variables que en un ejercicio de regresión lineal no se identificaron parámetros significativos

- **Escenario 3:** Tomando la base del escenario 1, Estandarizado las variables numéricas que presentan un rango numérico muy grande, es decir, "WeeklyWages", " HoursWorkedPerWeek", "DaysWorkedPerWeek", "UltimateIncurredClaimCost" y "days_Report".

  La estandarización centra los valores alrededor del cero así

$$Z = \frac{X - \mu}{\sigma} \tag{6}$$

  donde

  - $Z$: Es la variable estandarizada.
  - $X$: La variable a estandrizar.
  - $\mu$: El valor esperado de la variable X.
  - $\sigma$: La desviación estandar de X.

  Con esta transformación, Se detalla que siguen presentando los mismos outliers, sin embargo, el rango en el cual oscilan los datos, se ha acotado para las variables donde se realizó la estandarización.

- **Escenario 4:** Tomando la base del escenario 1, excluyendo los valores que se detectaron como outliers en el costo del siniestro, es decir, los valores de la variable "UltimateIncurredClaimCost" que estén fuera del rango

$$[1Q - 1{,}5 \times RIC, 3Q + 1{,}5 \times RIC] \tag{7}$$

  donde

  - $1Q$: Primer cuartil.
  - $3Q$: Tercer cuartil.
  - $RIC$: Rango intercuartílico= $3Q - 1Q$.

Una vez analizados los datos y hecho los ajustes pertinentes a la base, se entrenan los modelos de regresión seleccionados, con el fin de encontrar el que mejor se ajuste dados los escenarios

### 3.3.1. Regresión lineal

**Escenario 1**

del análisis de residuales del modelo resultante para el escenario 1 se puede notar:

- La imagen de residuales vs valores ajustados muestra que los puntos están agrupados en un lugar concreto, es decir, no presenta una notoria aleatoriedad, por tanto, la relación entre las variables independientes y las predictoras no presentan una relación lineal.
- En la gráfica "Normal Q-Q"se muestra que los residuos no presentan una distribución normal.
- La gráfica "Scale-Location"muestra que la varianza de los residuos no es constante en todo rango de observación, pues se ve como los datos presentan una tendencia a variar conforme sean mayores los valores ajustados.
- Se detalla que no existen puntos outliers que sesguen el modelo, pues en la gráfica Residuals vs Laverage"todos los puntos entán dentro del área delimitada por la distancia de Cook.

Dentro de los resultados del modelo lineal, también se presentan la significancia de las variables dentro del modelo descritas por el P-valor (**Pr(>|t|)**), adicionalmnete presenta los coeficientes de la regresión (**Estimate**). El resumen del modelo arroja los siguientes resultados:

Cuadro 4: Resumen modelo lineal: Escenario 1

| | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** | |
|---|---|---|---|---|---|
| **(Intercept)** | -1,05E+04 | 2,55E+03 | -4,135 | 3,56E-05 | *** |
| **Age** | 1,74E+02 | 1,76E+01 | 9,889 | <2,00e-16 | *** |
| **Gender.M** | -1,13E+03 | 4,41E+02 | -2,557 | 0,010573 | * |
| **MaritalStatus.M** | -3,75E+03 | 6,50E+02 | -5,763 | 8,38E-09 | *** |
| **MaritalStatus.S** | -4,26E+03 | 6,44E+02 | -6,617 | 3,74E-11 | *** |
| **DependentChildren** | 1,38E+03 | 3,57E+02 | 3,866 | 0,000111 | *** |
| **DependentsOther** | 7,19E+03 | 1,68E+03 | 4,279 | 1,89E-05 | *** |
| **WeeklyWages** | 1,45E+01 | 8,34E-01 | 17,372 | <2,00e-16 | *** |
| **PartTimeFullTime.F** | -3,07E+03 | 8,44E+02 | -3,641 | 0,000272 | *** |
| **HoursWorkedPerWeek** | -3,58E+01 | 3,69E+01 | -0,969 | 0,332758 | |
| **DaysWorkedPerWeek** | 1,94E+02 | 4,64E+02 | 0,419 | 0,675273 | |
| **A.OCURRENCIA** | 7,95E+02 | 3,65E+01 | 21,778 | <2,00e-16 | *** |
| **M.OCURRENCIA** | 3,43E+01 | 5,20E+01 | 0,661 | 0,50865 | |
| **days_Report** | 1,73E+01 | 2,92E+00 | 5,915 | 3,37E-09 | *** |

**Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Cuadro 4: Resumen modelo lineal: Escenario 1

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|

**Residual standard error: 28130 on 25270 degrees of freedom**
**Multiple R-squared: 0.06181, Adjusted R-squared: 0.06132**
**F-statistic: 128.1 on 13 and 25270 DF, p-value: <2.2e-16**

Dado el P- Valor obtenido en cada variable predictora, las variables más significativas del modelo lineal son las resaltadas con asteriscos (*) rojos en la tabla anterior, es decir, los que presentan P-valor más pequeño que son:

- La edad.
- Estado civil de quien reclamo.
- Número de hijos a cargo.
- Número de personas a cargo.
- Salario semanal.
- Jornada Laboral.
- Años de ocurrencia.
- Días transcurridos desde la ocurrencia hasta el aviso del siniestro.

Adicionalmente, dado el R-cuadrado ( 0.61 %), el modelo lineal no es un buen modelo para predecir el costo de las reclamaciones en este caso

### Escenario 2

Se validan los resultados de un nuevo modelo lineal ahora teniendo en cuenta las variables más significativas del escenario 1 y obteniendo lo siguiente.

Cuadro 5: Resumen modelo lineal: Escenario 2

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| **(Intercept)** | -1,09E+04 | 1,97E+03 | -5,502 | 3,80E-08 | *** |
| **Age** | 1,77E+02 | 1,75E+01 | 10,094 | <2,00e-16 | *** |
| **MaritalStatus.M** | -3,74E+03 | 6,50E+02 | -5,75 | 9,05E-09 | *** |
| **MaritalStatus.S** | -4,25E+03 | 6,44E+02 | -6,604 | 4,07E-11 | *** |
| **DependentChildren** | 1,37E+03 | 3,57E+02 | 3,844 | 0,000122 | *** |
| **DependentsOther** | 7,09E+03 | 1,68E+03 | 4,216 | 2,50E-05 | *** |
| **WeeklyWages** | 1,41E+01 | 7,99E-01 | 17,658 | <2,00e-16 | *** |
| **PartTimeFullTime.F** | -3,73E+03 | 6,29E+02 | -5,921 | 3,24E-09 | *** |

| | | | | | |
|---|---|---|---|---|---|
| **A.OCURRENCIA** | 7,99E+02 | 3,65E+01 | 21,905 | <2,00e-16 | *** |
| **days_Report** | 1,75E+01 | 2,92E+00 | 5,992 | 2,10E-09 | *** |

**Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 28140 on 25274 degrees of freedom**
**Multiple R-squared: 0.06149, Adjusted R-squared: 0.06116**
**F-statistic: 184 on 9 and 25274 DF, p-value: <2.2e-16**

Al extraer las variables menos significativas vemos que el modelo lineal sigue sin ser el más adecuado para predecir el costo del siniestro, pues el R cuadrado resultante es similar al escenario 1.

**Escenario 3**

Cuadro 6: Resumen modelo lineal: Escenario 3

| | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** | |
|---|---|---|---|---|---|
| **(Intercept)** | -0,5336223 | 0,0751289 | -7,103 | 1,26E-12 | *** |
| **Age** | 0,0061676 | 0,0006237 | 9,889 | <2,00e-16 | *** |
| **Gender.M** | -0,040005 | 0,0156472 | -2,557 | 0,010573 | * |
| **MaritalStatus.M** | -0,1328662 | 0,0230565 | -5,763 | 8,38E-09 | *** |
| **MaritalStatus.S** | -0,1511819 | 0,0228477 | -6,617 | 3,74E-11 | *** |
| **DependentChildren** | 0,0490093 | 0,0126758 | 3,866 | 0,000111 | *** |
| **DependentsOther** | 0,2550469 | 0,0596109 | 4,279 | 1,89E-05 | *** |
| **WeeklyWages** | 0,1252833 | 0,007212 | 17,372 | <2,00e-16 | *** |
| **PartTimeFullTime.F** | -0,1090055 | 0,0299381 | -3,641 | 0,000272 | *** |
| **HoursWorkedPerWeek** | -0,008861 | 0,0091483 | -0,969 | 0,332758 | |
| **DaysWorkedPerWeek** | 0,0037686 | 0,0089959 | 0,419 | 0,675273 | |
| **A.OCURRENCIA** | 0,0281827 | 0,0012941 | 21,778 | <2,00e-16 | *** |
| **M.OCURRENCIA** | 0,0012177 | 0,0018423 | 0,661 | 0,50865 | |
| **days_Report** | 0,0370181 | 0,0062587 | 5,915 | 3,37E-09 | *** |

**Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 0.9977 on 25270 degrees of freedom**
**Multiple R-squared: 0.06181, Adjusted R-squared: 0.06132**
**F-statistic: 128.1 on 13 and 25270 DF, p-value: <2.2e-16**

El modelo con este escenario presenta el mismo comportamiento que el escenario 1 y 2 a pesar de la estandarización realizada. Las variables independientes no presentan una relación lineal con el costo del siniestro.

**Escenario 4**

El análisis de residuales en este escenario presenta un comportamiento distinto en este escenario respecto a los demás, por lo que se tienen las siguientes observaciones:

- La imagen de residuales vs valores ajustados no muestran aleatoriedad, sin embargo, están un poco más cerca del 0 que los otros escenarios.
- En la gráfica "Normal Q-Q"se muestra que los residuos no presentan una distribución normal.
- La gráfica "Scale-Location"muestra que la varianza de los residuos no es constante en todo rango de observación, ya que se ve como los datos presentan una tendencia a variar cada vez menos conforme mayores sean los valores ajustados.
- Se detalla que no existen puntos outliers que sesguen el modelo, pues en la gráfica Residuals vs Laverage"todos los puntos están dentro del área delimitada por la distancia de Cook, pero en este caso son más lejanos a la línea limítrofe de la distancia de Cook.

Cuadro 7: Resumen modelo lineal: Escenario 4

| | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** | |
|---|---|---|---|---|---|
| **(Intercept)** | 1615,5973 | 380,8929 | 4,242 | 2,23E-05 | *** |
| **Age** | 40,4202 | 2,6261 | 15,392 | <2,00e-16 | *** |
| **Gender.M** | -753,9083 | 65,0093 | -11,597 | <2,00e-16 | *** |
| **MaritalStatus.M** | -1149,6528 | 98,175 | -11,71 | <2,00e-16 | *** |
| **MaritalStatus.S** | -1377,3937 | 97,2983 | -14,156 | <2,00e-16 | *** |
| **DependentChildren** | 77,4686 | 53,6739 | 1,443 | 0,148944 | |
| **DependentsOther** | 717,6315 | 252,3925 | 2,843 | 0,004469 | ** |
| **WeeklyWages** | 3,8208 | 0,1252 | 30,513 | <2,00e-16 | *** |
| **PartTimeFullTime.F** | -496,8551 | 128,1184 | -3,878 | 0,000106 | *** |
| **HoursWorkedPerWeek** | -13,8982 | 5,6262 | -2,47 | 0,013509 | * |
| **DaysWorkedPerWeek** | 82,232 | 69,4751 | 1,184 | 0,236577 | |
| **A.OCURRENCIA** | 109,4358 | 5,4274 | 20,164 | <2,00e-16 | *** |
| **M.OCURRENCIA** | 1,2519 | 7,6305 | 0,164 | 0,869687 | |
| **days_Report** | 1,5036 | 0,4624 | 3,251 | 0,00115 | ** |

**Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 3862 on 22106 degrees of freedom**
**Multiple R-squared: 0.1196, Adjusted R-squared: 0.1191**
**F-statistic: 231 on 13 and 22106 DF, p-value: <2.2e-16**

El modelo muestra un mejor ajuste que en otros escenarios; sin embargo, el R-cuadrado de 11,96 % sigue sin ser lo suficientemente alto para determinar que las variables tienen una relación lineal con la variable del costo del siniestro.

### 3.3.2. Arboles de decisión

El siguiente gráfico muestra los árboles construidos en cada uno de los escenarios predefinidos, se ilustra los nodos de decisión y el punto de inflexión en cada nodo.

**Escenario 1**

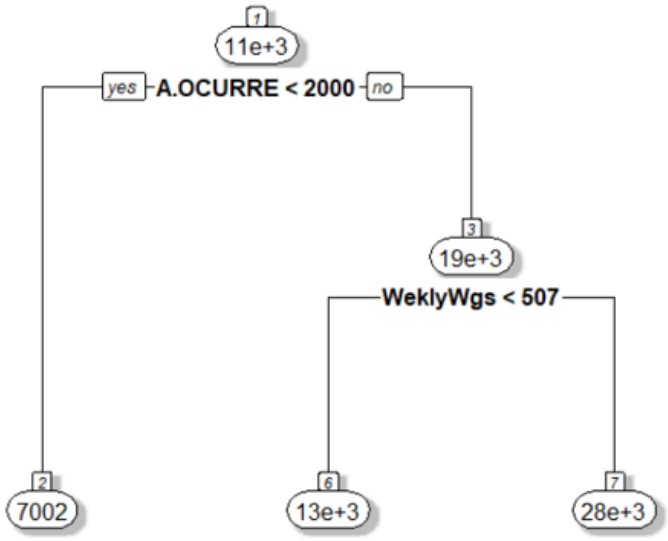


Figura 15: Árbol de regresión: Escenario 1

**Escenario 2**

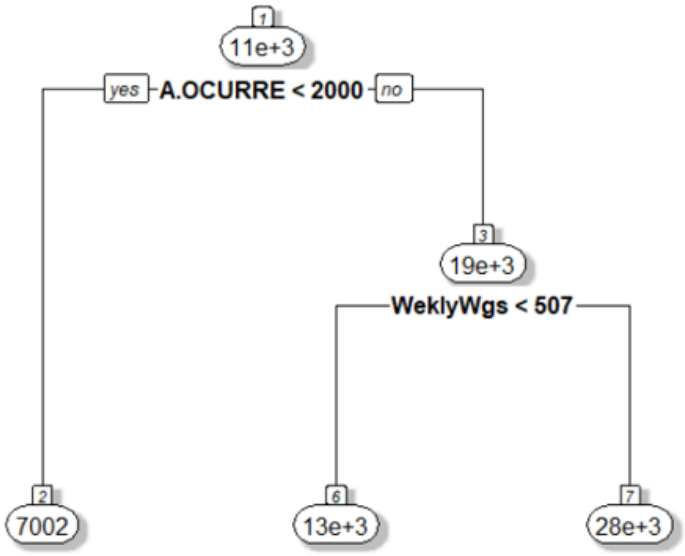


Figura 16: Árbol de regresión: Escenario 2

**Escenario 3**

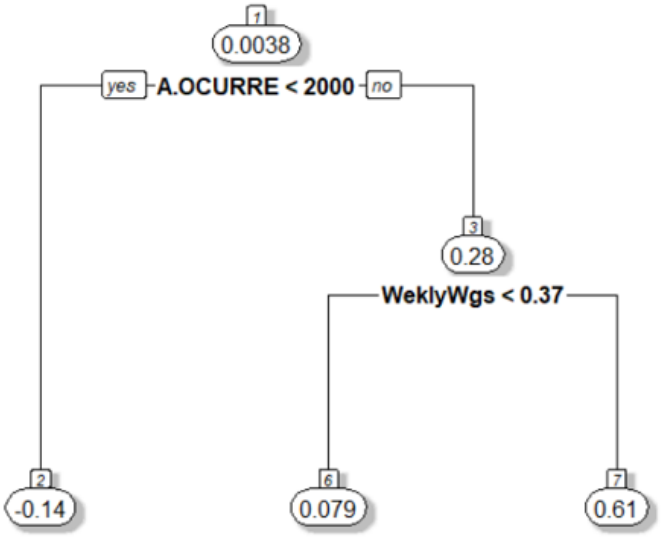


Figura 17: Árbol de regresión: Escenario 3

**Escenario 4**

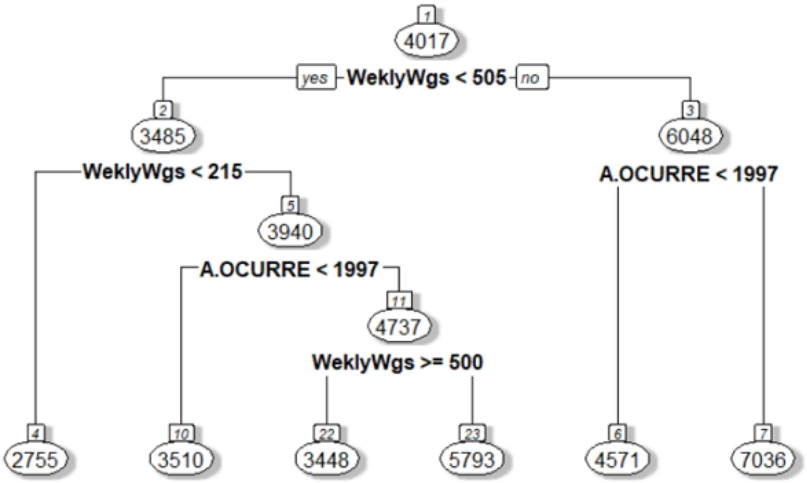


Figura 18: Árbol de regresión: Escenario 4

Los modelos resultantes para los escenarios 1 al 3 toman como referencia de decisión el año de ocurrencia y el salario semanal, especificando que el costo de las reclamaciones son diferentes antes y después del año 2000, en caso que el año de ocurrencia hubiese sido después del año 2000, compara el salario semanal de quien reclama, tomando como punto de comparación 507 unidades monetarias.

En el caso del escenario 4, donde se entrena el modelo sin tener en cuenta los outliers, el árbol generado es más robusto, tomando como referencia la variable año de ocurrencia y el salario semanal, pero en este caso, con puntos comparativos.

### 3.3.3. Red Neuronal

En cada uno de los gráficos de las redes neuronales se encuentra el número de nodos ocultos ($H_i$) y los errores asociados a cada nodo ($B_i$). En la parte izquierda de cada gráfico se muestra una tabla con el número de iteraciones realizadas hasta converger en cada uno de los escenarios propuestos.

## Escenario 1

| # weights: 211 | | | |
|---|---|---|---|
| initial | | value | 179.300,57 |
| iter | 10 | value | 111,32 |
| iter | 20 | value | 70,97 |
| iter | 30 | value | 47,23 |
| iter | 40 | value | 29,60 |
| iter | 50 | value | 28,00 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| iter | 970 | value | 26,02 |
| iter | 980 | value | 26,02 |
| iter | 990 | value | 26,02 |
| iter | 1000 | value | 26,02 |
| iter | 1010 | value | 26,02 |
| iter | 1020 | value | 26,02 |
| iter | 1030 | value | 26,02 |
| iter | 1040 | value | 26,02 |
| final | | value | 26,02 |
| converged | | | |

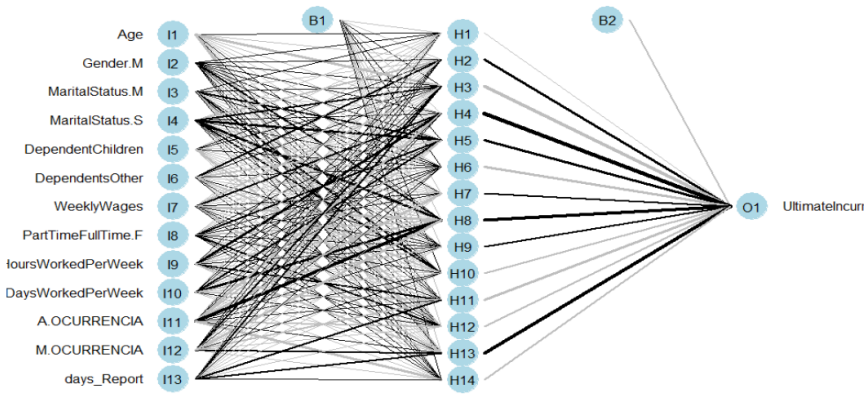


Figura 19: Red neuronal: Escenario 1

## Escenario 2

| # weights: 67 | | | |
|---|---|---|---|
| initial | | value | 56476 |
| iter | 10 | value | 29,588 |
| iter | 20 | value | 28,731 |
| iter | 30 | value | 28,378 |
| iter | 40 | value | 27,575 |
| iter | 50 | value | 27,352 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| iter | 220 | value | 26,688 |
| iter | 230 | value | 26,682 |
| iter | 240 | value | 26,679 |
| iter | 250 | value | 26,675 |
| iter | 260 | value | 26,675 |
| iter | 270 | value | 26,672 |
| iter | 280 | value | 26,668 |
| iter | 290 | value | 26,668 |
| final | | value | 26,668 |
| converged | | | |

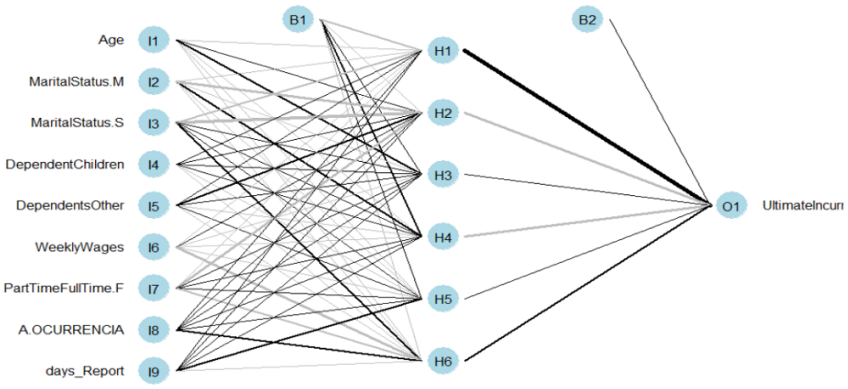


Figura 20: Red neuronal: Escenario 2

## Escenario 3

| # weights: 91 | | | |
|---|---|---|---|
| initial | | value | 26954 |
| iter | 10 | value | 26810 |
| iter | 20 | value | 26795 |
| iter | 30 | value | 26379 |
| iter | 40 | value | 25901 |
| iter | 50 | value | 25787 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| iter | 290 | value | 24850 |
| iter | 300 | value | 24840 |
| iter | 310 | value | 24835 |
| iter | 320 | value | 24834 |
| iter | 330 | value | 24832 |
| iter | 340 | value | 24830 |
| iter | 350 | value | 24817 |
| iter | 360 | value | 24815 |
| final | | value | 24815 |
| converged | | | |

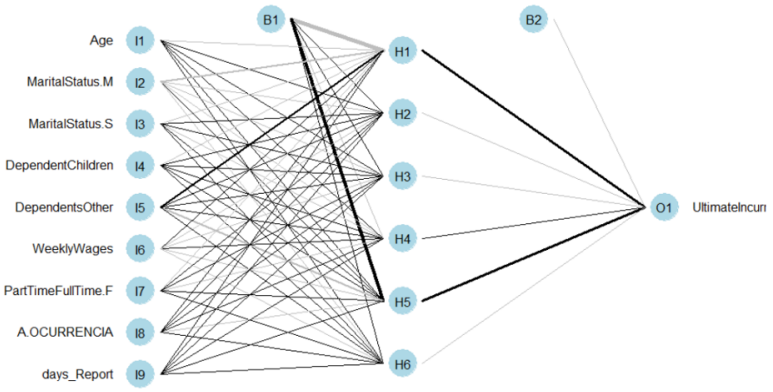


Figura 21: Red neuronal: Escenario 3

## Escenario 4

| # weights: 91 | | | |
|---|---|---|---|
| initial | | value | 138,826 |
| iter | 10 | value | 10,225 |
| iter | 20 | value | 8,516 |
| iter | 30 | value | 2,741 |
| iter | 40 | value | 0,508 |
| iter | 50 | value | 0,491 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| iter | 170 | value | 0,441 |
| iter | 180 | value | 0,441 |
| iter | 190 | value | 0,441 |
| iter | 200 | value | 0,441 |
| iter | 210 | value | 0,440 |
| iter | 220 | value | 0,440 |
| iter | 230 | value | 0,440 |
| iter | 240 | value | 0,440 |
| final | | value | 0,440 |
| converged | | | |

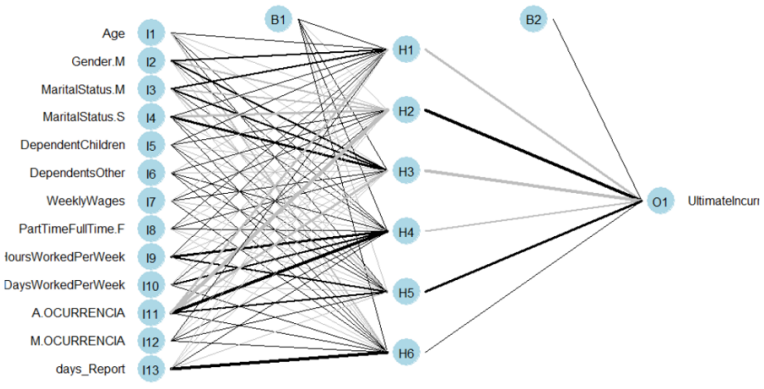


Figura 22: Red neuronal: Escenario 4

En el primer escenario se muestra que la red neuronal asoció 211 pesos para determinar el valor del siniestro con un número de 14 nodos en una de las capas internas, el modelo encontró convergencia en la iteración número 1.040. Lo que evidencia las redes neuronales de los escenarios 2,3 y 4 es que a comparación del escenario 1, fueron necesarias menos iteraciones para converger, y un número de 6 nodos en las capas internas, esto quiere decir, que los datos del escenario 1 no muestran una relación tan "sencilla"de comprender entre las variables independientes y el costo del siniestro.

### 3.3.4. Bagging

En cada una de las gráficas a continuación se muestra la significancia de las variables en cada escenario propuesto a la hora de ajustar el modelo Bagging:

**Escenario 1**

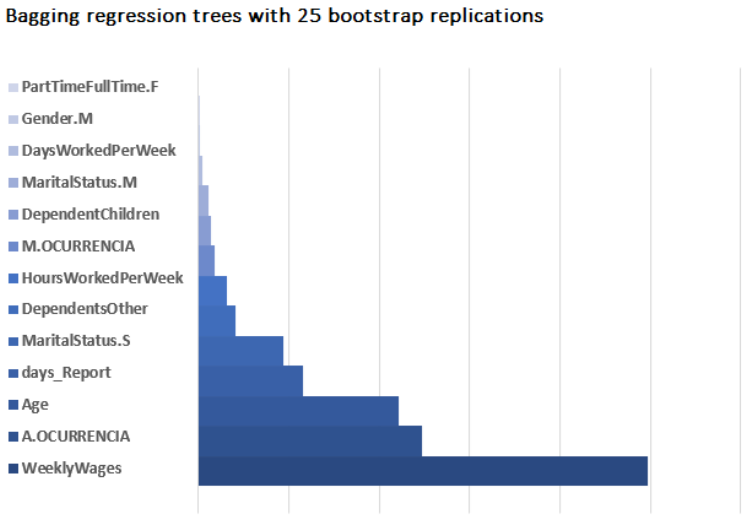


Figura 23: Variables significativas Bagging: Escenario 1

**Escenario 2**

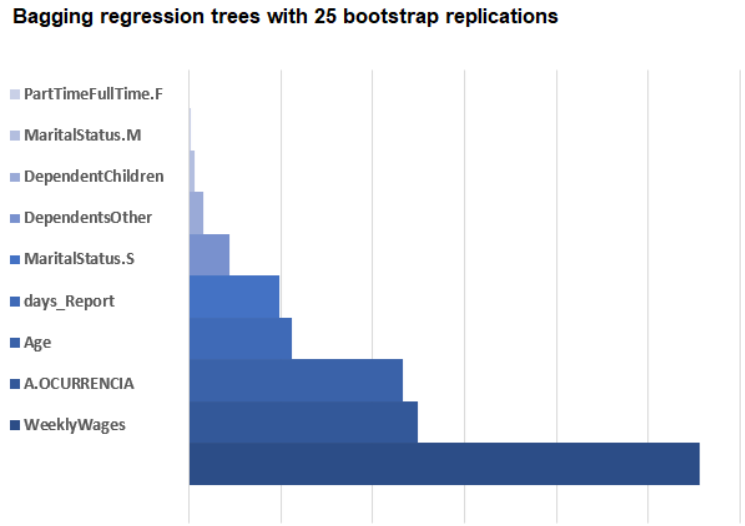


Figura 24: Variables significativas Bagging: Escenario 2

## Escenario 3

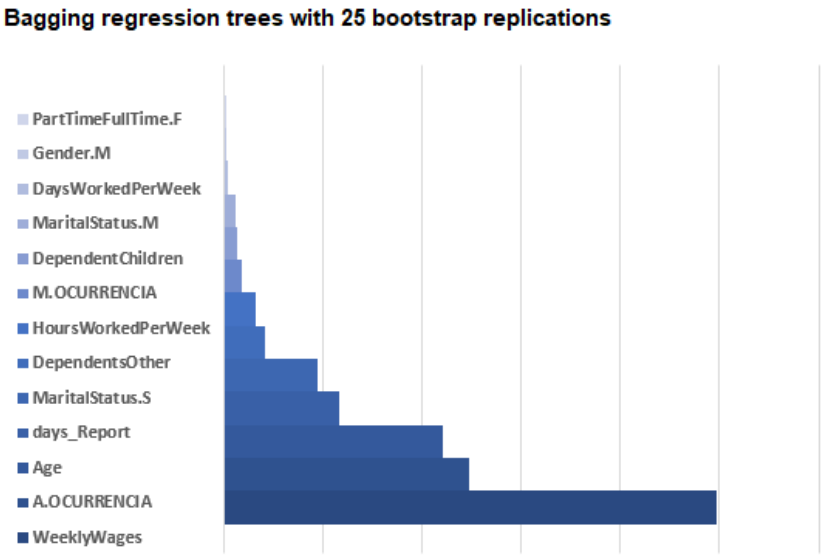


Figura 25: Variables significativas Bagging: Escenario 3

## Escenario 4

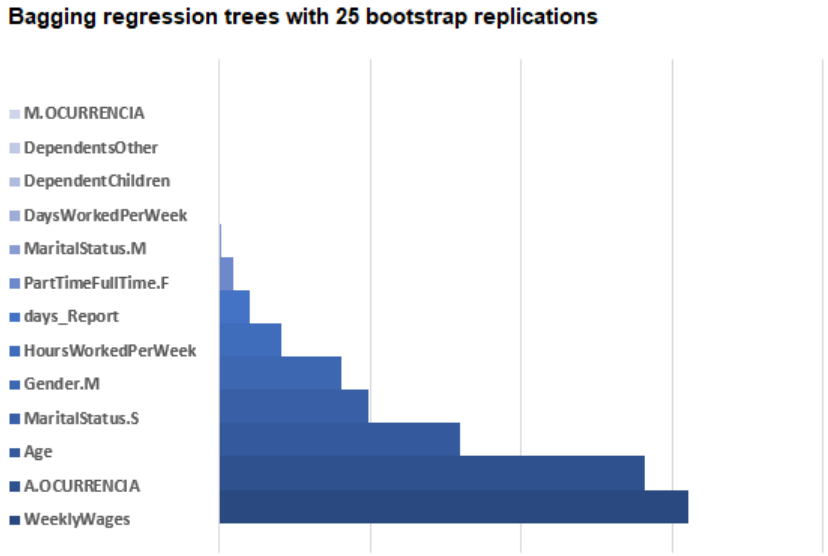


Figura 26: Variables significativas Bagging: Escenario 4

Para cada uno de los escenarios, la variable que representa el salario semanal es la más significativa.

### 3.3.5. Gradient Boosting

En cada una de las gráficas, a continuación se muestra la significancia de las variables en cada escenario propuesto a la hora de ajustar el modelo Boosting:

**Escenario 1**

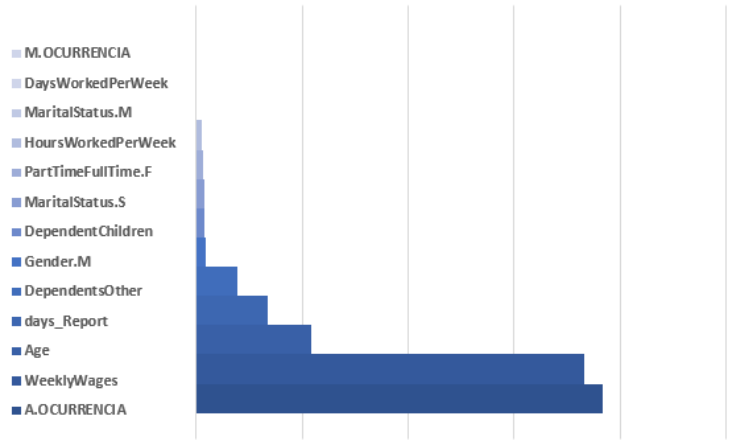


Figura 27: Variables significativas Boosting: Escenario 1

**Escenario 2**

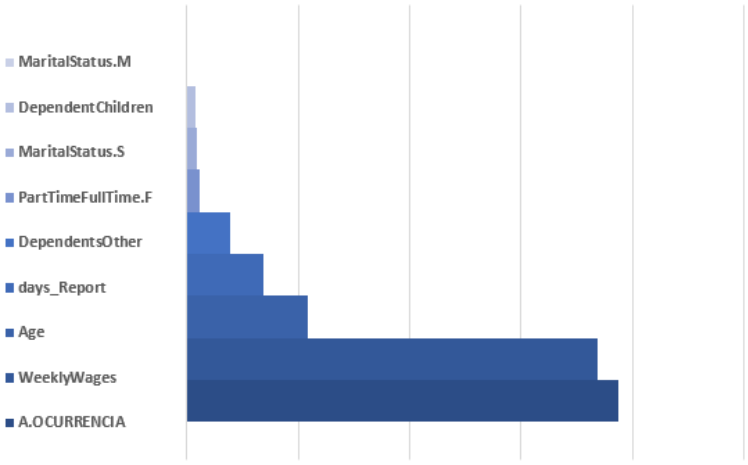


Figura 28: Variables significativas Boosting: Escenario 2

**Escenario 3**

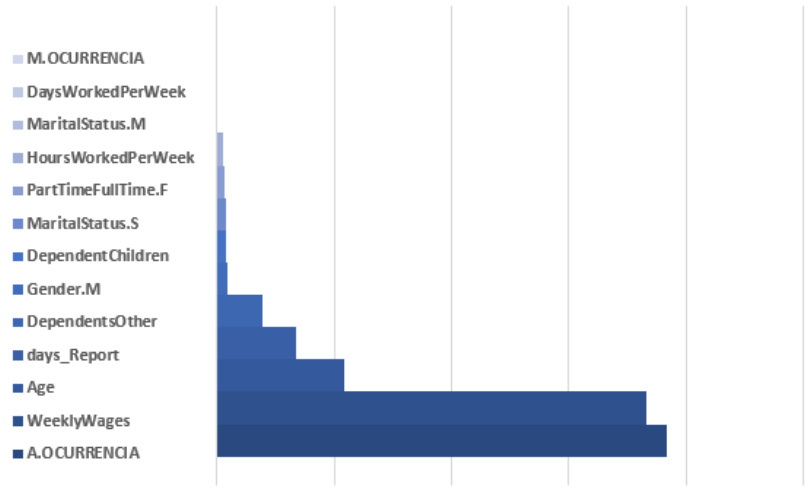


Figura 29: Variables significativas Boosting: Escenario 3

**Escenario 4**

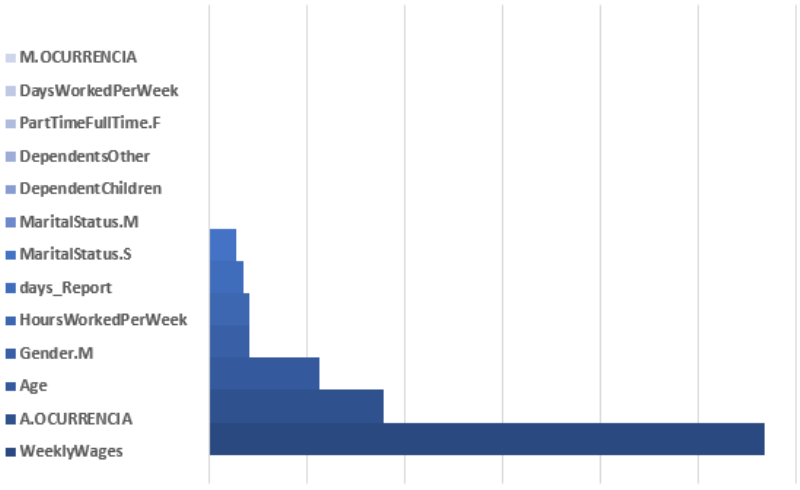


Figura 30: Variables significativas Boosting: Escenario 4

A comparación de los modelos resultantes con Bagging, la variable más significativa de los modelos Boosting es la que representa el año de ocurrencia seguido del salario semanal, excepto para el escenario 4 donde se invierte el resultado, y en este escenario, el nivel de significancia entre el año de ocurrencia y el salario semanal son notoriamente diferentes.

## 3.4. Resultados

Una vez entrenados los modelos con la base de entrenamiento, se toma el conjunto de validación con el fin de ejecutar cada uno de los modelos y comparar los resultados de predicción con el error absoluto medio (4) y el R-cuadrado (5) en cada uno de los escenarios y determinando cuál modelo se ajusta mejor a predecir el costo de las reclamaciones.

Cuadro 8: Error absoluto Medio

| EAM | | | | | |
|---|---|---|---|---|---|
| **Escenario** | **Regresión lineal** | **Árboles de Regresión** | **Red Neuronal** | **Bagging** | **Gradient Boosting** |
| **1** | 11210,71 | 11105,39 | 11161,46 | 11058,36 | 10918,07 |
| **2** | 11223,50 | 11.105,39 | 11189,06 | 11059,73 | 10918,69 |
| **3** | 11210,71 | 11105,39 | 10927,41 | 11058,36 | 10918,07 |
| **4** | 3010,46 | 3004,98 | 2981,26 | 3008,50 | 2971,94 |

De los resultados comparativos entre modelos y escenarios podemos inferir que:

- En general, ninguno modelo presenta un porcentaje significativo de ajuste con los datos para determinar el costo de las reclamaciones, dentro de los escenarios tratados.
- Los modelos resultantes presentan un mejor ajuste al extraer los outliers de la base original (Escenario 4), en comparación de los demás escenarios.
- El que modelo que mejore se ajusta a cada escenario es Gradiente Boosting seguido de Bagging por lo que se puede concluir que los modelos de ensamble tienen mejor capacidad de predicción que los demás modelos probados.
- Se corrobora que al extraer las variables poco significativas o estandarizar la base para el modelo de árboles de regresión no tienen relevancia, pues como vimos anteriormente en las gráficas de los modelos resultantes, las variables extraídas en el escenario 2 no pertenecían a los nodos de decisión y acotar las variables tampoco marcó diferencia, por eso se evidencia el mismo error cuadrático medio en los 3 primeros escenarios.
- Para este caso particular, el comportamiento de las reclamaciones no mostró un escenario idóneo para la aplicación de modelos de regresión supervisada dado el $R^2$ presente en los resultados.

# 4. Conclusiones

Al detallar los resultados obtenidos del ejercicio aplicativo, se presenta que existen modelos con mejor ajuste que el modelo lineal, como en este caso lo fue Gradient Boosting. Adicionalmente, es útil conocer modelos que expliquen comportamientos de variables aleatorias, en este caso siniestros, sin alguna distribución específica como los modelos derivados de los árboles de decisión o redes neuronales, puesto que en la realidad varias veces se presenta esta situación.

En este caso particular, se presenta la importancia de análisis exploratorio de datos o preprocesamiento de datos, ya que genera una noción amplia a la hora de definir que modelos pueden presentar mejor ajuste a los datos de estudio.

A pesar de los resultados poco favorables para la base estudiada, las predicciones con Machine Learning pueden presentar una gran relevancia en el estudio actuarial, concretamente, en la siniestralidad. Aportando nuevas técnicas que pueden ayudar a mejorar el ajuste de las proyecciones realizadas, por ejemplo, a la hora de estimar el costo de las reclamaciones futuras, que hace parte de una de las tareas a ejecutar para el cálculo de la tasa pura de riesgo. Adicionalmente, implementar técnicas de Machine Learning incentiva a analizar el comportamiento de los datos con mayor detalle previo a entrenar un modelo, permitiendo justificar los pronósticos obtenidos.

Se evidencia la necesidad del actuario de encontrar las relaciones en el comportamiento de las reclamaciones según las características del mismo, en esa labor, se determinan anomalías en el comportamiento siniestral, o bien incidencias e irregularidades que pueden estar asociadas a una mala recopilación de información. Esto genera un beneficio no solo para realizar un cálculo sin sesgos, sino, a nivel compañía, identificar la procedencia del error y evitar que se repita.

Cabe resaltar que la autora del presente trabajo realizó el mismo ejercicio práctico aplicado a una base de datos relacionada con los recargos realizados a las primas de las pólizas de acuerdo a las características del asegurado, presentando mejores resultados de ajuste ($R^2$ superior al 80 %), sugiriendo que la escogencia de la base de datos de reclamos por accidentes laborales en esta ocasión no fue una buena fuente para aplicar los modelos de Machine Learning.

Como continuación de este trabajo se pueden explorar otros ámbitos relacionados a las funciones de un actuario en las que se pueden aplicar las técnicas que aporta el Machine Learning, como la segmentación de clientes para mejorar la persistencia en seguros y también ofrecer una tarifa más adecuada según el perfil del asegurado y la prevención y detección de fraude en las reclamaciones, mejorando el comportamiento siniestral.

Adicionalmente, dada la utilidad e importancia que presenta el análisis exploratorio de datos a la hora de querer aplicar algún modelo de regresión, se pensaría la oportunidad de desarrollar una herramienta que proporcione información gráfica inmediata una vez se ingrese la información esto con el fin de disminuir el tiempo en el preprocesamiento de datos, esta misma plataforma se puede utilizar para la una tarea frecuente del actuario como lo es determinar las características del comportamiento tanto en si-

niestros como en asegurados de una póliza. En la herramienta de programación "R" de uso gratuito se puede llevar a cabo mediante el paquete Shiny que facilita la creación de aplicaciones de interfaz interactivas.

# 5. Anexos

Se anexa un documento adicional en formato HTML denominado "Código.html" que contiene el código en R desarrollado y la ejecución de cada uno de los comandos. Este archivo soporta los resultados presentes en el trabajo.

# Referencias


[Álvarez, S. (2018)] Álvarez, S. (2018) *Análisis del Big Data en los Seguros: Modelos Predictivos* (tesis de maestría). Máster universitario en ciencas actuariales y Financieras. Universitat de León.

[Bastien, M. ] Bastien, M. *The Pricing of Group Life Insurance Schemes.* Recuperado de http://www.actuariesindia.org/subMenu.aspx?id=11&val=Downloads en Febrero 2021.

[Blier-Wong, C. (2020) ] Blier-Wong, C.; Cossette, H.;Lamontagne, L.; Marceau, E. (2020) *Machine Learning in P&C Insurance: A Review for Pricingand Reserving.* Recuperado de https://www.mdpi.com/2227-9091/9/1/4 en ocubre de 2021.

[Fortune. (2022)] Fortune Business Insights, (2022) *Machine Learning (ML) market Size, Share & COVID-19 Impact Analysis* Recuperado de https://www.fortunebusinessinsights.com/machine-learning-market-102226 el 20 de marzo de 2022.

[Guillen, M. (2018)] Guillen, M. & Pesantez, J. (2018) *Machine Learning Y Modelización Predictiva Para La Tarificación En El Seguro De Automóviles.* Revista Anales del Instituto de Actuarios Españoles.

[Insurance. (2021)] Insurance Innovation Reporter, (2021) *AI's Tremendous Potential for P&C Insurance: SMA Study* Recuperado de https://iireporter.com/ais-tremendous-potential-for-pc-insurance-sma-study/ en abril de 2021.

[Kunce, J. (2017)] Kunce, J. & Chatterjee, S. (2017) *A MACHINE-LEARNING APPROACH TO PARAMETER ESTIMATION* Arlington, Virginia.: Editorial Casualty Actuarial Society.

[Miller, D. (2017)] Miller, D. (2017) *Statistics for Data Science: Leverage the power of statistics for Data Analysis,Classification, Regression, Machine Learning, and Neural Networks.*: Editorial Packt Publishing Ltd.

[Oscar, P. (2017)] Oscar, P. (2017) *Big Data y el sector asegurador* (tesis de maestría). Máster en Dirección de Entidades Aseguradoras y Financieras. Universitat de Barcelona.

[Padilla, A. (2017)] Padilla, A., Guillen, M. & Bolancé, C. (2017) *Big-Data Analytics En Seguros.* Revista Anales del Instituto de Actuarios Españoles.

[Perkins, S. (2020)] Perkins, S. ; Davis, H. ; & Valerie du Preez *Practical Data Science for Actuarial Tasks: A practical example of data science considerations.* Institute and Faculty of Actuaries.

[QuantumBlack. (2021)] QuantumBlack, AI by McKinsey, (2021) *The state of AI in 2021* Recuperado de https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021 en marzo de 2022.

[Ramasubramanian, K. (2019)] Ramasubramanian, K. & Singh, A. (2019) *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R.* New Delhi, India: Editorial Apress.

[Spedicato, G. (2018) ] Spedicato, G., Dutang , C. & Petrini. L. (2018) *achine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs.* Recuperado de https://hal.archives-ouvertes.fr/hal-01942038 en Febrero 2021.