**Maestría en Ciencias Actuariales**

# Determinación del impacto de incluir variables telemáticas en la construcción de modelos de prima pura para pólizas de auto en Colombia

**Diego Alejandro Espitia Villalobos**

**Bogotá, D.C., 15 de mayo de 2023**

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
VIGILADA MINEDUCACIÓN

# Determinación del impacto de incluir variables telemáticas en la construcción de modelos de prima pura para pólizas de auto en Colombia

**Tesis para optar al título de Magíster en Ciencias Actuariales**

**Andres Felipe Julio Niño**

**Director**

**Bogotá, D.C., 15 de mayo de 2023**

La tesis de maestría titulada “Determinación del impacto de incluir variables telemáticas en la construcción de modelos de prima pura para pólizas de auto en Colombia”, presentada por Diego Alejandro Espitia Villalobos, cumple con los requisitos establecidos para optar al título de Magíster en Ciencias Actuariales.

Director de la tesis

Andrés Felipe Julio Niño

Jurado

Nombre 2

Jurado

Nombre 3

Bogotá, D.C., día de mes de año (fecha de aceptación del trabajo por parte del jurado)

## Dedicatoria

A Susi, Mathi, mis padres y Marino quienes, con su paciencia, cariño, apoyo, ánimo y profunda comprensión, me motivan a ser cada día mejor.

## Agradecimientos

Mi más sincero agradecimiento a mi tutor, Andrés Felipe Julio Niño, por su tiempo, asesoría y orientación en el desarrollo de este proyecto. A la comunidad de la Maestría en Ciencias Actuariales de la Escuela Colombiana de ingeniería Julio Garavito, desde las directivas hasta mis compañeros de estudio, por todo su apoyo, comentarios y ayuda a lo largo de esta etapa.

Finalmente, un profundo agradecimiento a todas aquellas personas que, de una u otra manera, contribuyeron con su apoyo, consejos y palabras de aliento en la realización de este trabajo.

## Resumen

El presente trabajo tiene como objetivo determinar el efecto que supone el incluir variables telemáticas provenientes de dispositivos IoT en la construcción de modelos de prima pura de riesgo. Debido a la inexistencia de bases de datos reales con variables telemáticas en Colombia, se utilizó una base de datos sintética generada a partir del artículo “Synthetic Dataset Generation of Driver Telematics” (So, Boucher, & Valdez, 2021).

Para cumplir con el objetivo, se construyó un modelo de tarifación con variables tradicionales (género, edad, estado civil, entre otras) como punto de referencia para comparar los demás modelos construidos. Luego, se crearon diferentes modelos de tarifación utilizando modelos GLM con diferentes combinaciones de variables tradicionales y telemáticas. La comparación de los modelos se realizó a partir del estadístico AIC, que permite seleccionar el modelo con mejor ajuste a los datos. (Goldburd, Khare, Tevet, & Guller, 2020)

Los resultados mostraron que la inclusión de variables telemáticas en los modelos de frecuencia mejora el AIC, aunque no significativamente. En particular, las variables telemáticas de aceleración y frenado resultaron ser las más importantes al ser estadísticamente significativas en la mayoría de los diferentes modelos construidos. En contraste, los modelos de severidad y de prima pura (Tweedie) no lograron ajustar con la misma eficiencia que los modelos de frecuencia.

Entre las limitaciones del estudio se encontró la falta de datos reales con variables telemáticas en Colombia, lo que limitó la capacidad de generalización de los resultados. Además, se utilizaron modelos GLM tradicionales para la construcción de los modelos de frecuencia, severidad y prima pura en lugar de metodologías modernas de modelación como RF o XGBoost.

En conclusión, los resultados sugieren que el uso de datos telemáticos provenientes de dispositivos IoT puede mejorar la capacidad predictiva de los modelos de tarifación de pólizas de autos individuales en Colombia principalmente para explicar el componente de frecuencia de los reclamos, aunque estos resultados se deben revisar muy bien antes de hacer generalizaciones y, en lo posible, replicar los resultados con datos no simulados que permitan la construcción de modelos base y así cómo de modelos con variables telemáticas más robustos y que proporcionen resultados más confiables.

## Índice general


Resumen ..... 6

Índice general ..... 7

Índice de tablas ..... 9

Índice de figuras ..... 10

Índice de anexos ..... 12

1. Introducción ..... 13
2. Estado del arte ..... 13
3. Problema ..... 17
   - 3.1. Planteamiento del problema ..... 17
   - 3.2. Árbol de problemas ..... 18
   - 3.3. Árbol de objetivos ..... 18
   - 3.4. Objetivos ..... 19
     - 3.4.1. Objetivo general ..... 19
     - 3.4.2. Objetivos específicos ..... 19
   - 3.5. Alcances y limitaciones ..... 19
     - 3.5.1. Alcance ..... 19
     - 3.5.2. Limitaciones ..... 20
4. Metodología ..... 20
   - 4.1. Definición de la metodología de trabajo ..... 20
   - 4.2. Adquisición de datos ..... 21
   - 4.3. Tratamiento de las bases reales ..... 24
   - 4.4. Criterios de inclusión y selección de variables ..... 25
   - 4.5. Unión de las bases real y simulada ..... 26
   - 4.6. Alineación de la base construida con la base real ..... 27
   - 4.7. Selección de variables ..... 27
   - 4.8. Segmentación de variables ..... 36
   - 4.9. Construcción de modelos ..... 42

4.9.1. Modelos de frecuencia: 45
4.9.2. Modelos de severidad: 46
4.9.3. Modelos Tweedie: 48
4.10. Resultados y comparación de los modelos 49
4.10.1. Modelos de frecuencia 50
4.10.2. Modelos de severidad 50
4.10.3. Modelos Tweedie 51
5. Conclusiones y recomendaciones 52
6. Referencias bibliográficas 54

## Índice de tablas

Tabla 1. Campos de la base de pólizas reales ..... 21
Tabla 2. Campos de la base de siniestros reales ..... 22
Tabla 3. Campos incluidas en la base de datos simulada ..... 23
Tabla 4. Factores de desarrollo de pólizas reales ..... 25
Tabla 5.Conteo de registros por número de reclamos ..... 27
Tabla 6. Conteo de registros por número de reclamos de la base unificada ..... 27
Tabla 7. AIC de los modelos de frecuencia. Elaboración propia ..... 50
Tabla 8. AIC de los modelos de severidad. Elaboración propia ..... 50
Tabla 9. AIC de los modelos Tweedie. Elaboración propia ..... 51

## Índice de figuras

Figura 1. Árbol de problemas ..... 18
Figura 2. Árbol de objetivos ..... 18
Figura 3. Campos de la base unificada ..... 28
Figura 4. Correlación entre las variables telemáticas. Elaboración propia ..... 29
Figura 5. Variables telemáticas con correlación superior a 0.7. Elaboración propia ..... 30
Figura 6. Correlación de las variables Pct.drive. Elaboración propia ..... 31
Figura 7. Contribución de las variables Pct.drive a la dimensión 1 del PCA ..... 32
Figura 8. Correlación de las variables Accel. Elaboración propia ..... 32
Figura 9. Contribución de las variables Accel a la dimensión 1 del PCA ..... 33
Figura 10. Correlación de las variables Brake. Elaboración propia ..... 33
Figura 11. Contribución de las variables Brake a la dimensión 1 del PCA. Elaboración propia ..... 34
Figura 12. Correlación de las variables Right. Elaboración propia ..... 34
Figura 13. Contribución de las variables Right a la dimensión 1 del PCA. Elaboración propia ..... 35
Figura 14. Correlación de las variables Left. Elaboración propia ..... 35
Figura 15. Contribución de las variables Left a la dimensión 1 del PCA. Elaboración propia ..... 36
Figura 16. Tasa de reclamos por categorías de Insured.age. Elaboración propia ..... 37
Figura 17. Tasa de reclamos por categoría de Car.age. Elaboración propia ..... 38
Figura 18. Tasa de reclamos por categoría de Years.noclaims. Elaboración propia ..... 39
Figura 19. Tasa de reclamos por categoría de Pct.drive.wkday. Elaboración propia ..... 39
Figura 20. Tasa de reclamos por categoría de Accel.09miles. Elaboración propia ..... 40
Figura 21. Tasa de reclamos por categoría de Accel.11miles. Elaboración propia ..... 40
Figura 22. Tasa de reclamos por categoría de Brake.09miles. Elaboración propia ..... 41
Figura 23. Tasa de reclamos por categoría de Brake.11miles. Elaboración propia ..... 41
Figura 24. Tasa de reclamos por categoría de Left.turn.intensity10. Elaboración propia ..... 42
Figura 25. Tasa de reclamos por categoría de Right.turn.intensity10. Elaboración propia 42
Figura 27. Significancia estadística de los parámetros del modelo de frecuencia IoT 1. Elaboración propia ..... 46
Figura 36. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Backward". Elaboración propia ..... 48
Figura 41. Significancia estadística de los parámetros del modelo Tweedie usando StepAIC en dirección Backward. Elaboración propia ..... 49
Figura 42. Factores de desarrollo de pólizas reales ..... 56
Figura 26. Significancia estadística de los parámetros del modelo de frecuencia base. Elaboración propia ..... 57
Figura 27. Significancia estadística de los parámetros del modelo de frecuencia IoT 1. Elaboración propia ..... 58
Figura 28. Significancia estadística de los parámetros del modelo de frecuencia IoT 2. Elaboración propia ..... 59

Figura 29. Significancia estadística de los parámetros del modelo de frecuencia IoT 3 usando StepAIC en dirección "Both". Elaboración propia ..... 60
Figura 30. Significancia estadística de los parámetros del modelo de frecuencia IoT 3 usando StepAIC en dirección "Backward". Elaboración propia ..... 61
Figura 31. Significancia estadística de los parámetros del modelo de frecuencia IoT 3 usando StepAIC en dirección "Forward". Elaboración propia..... 62
Figura 32. Significancia estadística de los parámetros del modelo de severidad base. Elaboración propia..... 63
Figura 33. Significancia estadística de los parámetros del modelo de severidad IoT 1. Elaboración propia..... 64
Figura 34. Significancia estadística de los parámetros del modelo de severidad IoT 2. Elaboración propia..... 65
Figura 35. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Both". Elaboración propia ..... 66
Figura 36. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Backward". Elaboración propia ..... 67
Figura 37. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Forward". Elaboración propia..... 67
Figura 38. Significancia estadística de los parámetros del modelo Tweedie base. Elaboración propia..... 69
Figura 39. Significancia estadística de los parámetros del modelo Tweedie IoT 1. Elaboración propia..... 70
Figura 40. Significancia estadística de los parámetros del modelo Tweedie IoT 2. Elaboración propia..... 71
Figura 41. Significancia estadística de los parámetros del modelo Tweedie usando StepAIC en dirección Backward. Elaboración propia ..... 72

## Índice de anexos

Anexo 1 Factores de desarrollo de pólizas reales ..... 56
Anexo 2 Resultados de los modelos de frecuencia ..... 57
Anexo 3 Resultados de los modelos de severidad ..... 63
Anexo 4 Resultados de los modelos de Tweedie ..... 69

## 1. Introducción

El Internet de las cosas (IoT) se ha convertido en una de las innovaciones más disruptivas de la última década. Los dispositivos IoT están presentes en prácticamente todos los aspectos de nuestras vidas, transformando la manera en que interactuamos con el mundo y generando una cantidad inmensa de datos que utilizan las diferentes industrias para conocer nuestros comportamientos, rutinas, entre otras, lo que les ha permitido generar productos más acordes a las necesidades de sus usuarios.

El sector asegurador no es ajeno a esta tendencia y está siendo profundamente impactado por los dispositivos IoT, ya que estos ofrecen nuevas oportunidades para mejorar la gestión de riesgos y reducir costos en un mercado altamente competitivo. Los productos de seguro basados en IoT están transformando la forma en que se miden y gestionan los riesgos, lo que permite a las aseguradoras personalizar las tarifas y mejorar la experiencia del cliente. En particular, los productos de seguros para autos están empezando a utilizar dispositivos que se instalan en los vehículos y recopilan datos en tiempo real sobre la forma en que se conduce.

Por lo anterior, se plantea como el principal objetivo del presente trabajo realizar el análisis del impacto de los dispositivos IoT en el sector asegurador específicamente para el producto de seguro de autos e identificar cómo pueden impactar en la construcción de tarifas más acordes a los hábitos de conducción de los clientes.

Para ello, se describirá la metodología de adquisición de los datos, tanto reales como simulados, la generación de una base única de trabajo, la construcción de una tarifa base de comparación y la generación de tarifas que tengan en cuenta variables suministradas por dispositivos IoT. Finalmente se desarrollarán las conclusiones del trabajo, y recomendaciones generales para la mejora continua de los resultados encontrados.

## 2. Estado del arte

El término “Internet de las Cosas” (IoT) fue empleado por primera vez en 1999 por el pionero británico Kevin Ashton para describir un sistema en el cual los objetos del mundo físico se podían conectar a Internet por medio de sensores (Rose, Eldrige, & Chapin, 2015), permitiendo interactuar entre los seres humanos y su entorno de formas que antes hubiesen sido inimaginables. Para dar un ejemplo, hoy día existen “neveras inteligentes” que informan a sus dueños cuando las provisiones empiezan a escasear. Otro ejemplo son los dispositivos de comunicación desarrollados por Google (Google Home) (Google, 2021) o Amazon (Alexa) (Amazon, 2021) que permiten controlar diferentes elementos del hogar como televisores, equipos de sonido, luces, etc (siempre y cuando sean compatibles con dicha tecnología).

Una de las características más importantes de los dispositivos IoT es su capacidad de generar grandes cantidades de datos acerca del uso que se les da. Por ejemplo, los relojes inteligentes (smartwatches) o las bandas inteligentes (smartbands) monitorean diferentes

aspectos de la actividad física diaria de una persona como el número de pasos realizados al día, el tiempo de reposo y la calidad de este o la cantidad de kilómetros recorridos en un paseo realizado en bicicleta, así como la ruta elegida (Deloitte, s.f.). Producto de estos datos, es posible calcular la cantidad de calorías quemadas en una sesión de ejercicio, el nivel de oxígeno en la sangre y muchas otras medidas que permiten tener una idea general del estado de salud de una persona, sin la necesidad de recurrir constantemente a un especialista de la salud.

Así mismo, desde hace varios años se ha desarrollado un nuevo paradigma de procesamiento de información denominado Big Data el cual consta de una serie de herramientas enfocadas en la integración, procesamiento y visualización de grandes volúmenes de información provenientes de distintas fuentes estructuradas (hojas de cálculo, bases de datos) y no estructuradas (videos, audio) a una gran velocidad. Esta capacidad de procesamiento ha sido principalmente relevante para el desarrollo y auge de los dispositivos IoT dando la capacidad de analizar, incluso en tiempo real, la información generada por estos dispositivos y así mismo, tener la capacidad de tomar decisiones basados en la misma (Oracle, s.f.).

De otro lado, el sector asegurador es un sector que tradicionalmente ha requerido la información del comportamiento de sus clientes para el desarrollo y construcción de modelos que permitan fijar el precio de sus productos, tanto por procesos propios como el análisis de patrones de comportamiento como para proyectar posibles siniestros que puedan tener sus asegurados, determinar un precio adecuado de sus productos teniendo en cuenta el riesgo asumido o cambiar a un enfoque de prevención sin dejar de lado la protección. Por su parte, por la regulación del sector exige una ventana de información histórica estadísticamente suficiente para realizar una adecuada fijación de precios.

Sin embargo, los modelos de tarifación desarrollados por las aseguradoras son generalmente construidos a partir de variables tradicionales como información geográfica, socioeconómica entre otras que, si bien, han sido útiles para el propósito definido, también dan un margen de mejoramiento mediante el uso de información no tradicional, como los datos de las redes sociales o de dispositivos de IoT.

En este sentido, diferentes compañías aseguradoras han desarrollado pilotos en los que han utilizado información proveniente de dispositivos IoT con el fin de calcular de forma más precisa el riesgo de un contrato puntual y de esta forma fijar un costo más preciso y competitivo para cada negocio (Deloitte, s.f.). Por ejemplo (Internet of Business, s.f.):

- Liberty Mutual

Liberty Mutual realizó una alianza con Google, de modo que los clientes que adquieran la póliza de hogar tengan un descuento en la adquisición del dispositivo Google Nest. Este equipo es un parlante que cuenta con una inteligencia artificial la cual permite interactuar, a través de comandos de voz, con diferentes dispositivos o aplicaciones, por ejemplo: a

través de un comando de voz se puede reproducir música desde Spotify en el dispositivo. De igual forma, el dispositivo informa o responde a las órdenes del propietario mediante audios o notificaciones al smartwatch o al smartphone.

En este sentido, Google Nest puede interactuar con alarmas de humo compatibles e informar al propietario, en tiempo real, sobre la detección de humo que pueda derivar en un incendio en la propiedad y además indicar la zona de su origen.

Contar con este dispositivo le ha permitido reducir las reclamaciones a esta aseguradora e incluso reducir el costo de la prima del producto.

- John Hancock

La aseguradora John Hancock fue la primera entidad en aprovechar las ventajas de los dispositivos IoT para el monitoreo del bienestar físico de sus asegurados. A través de una alianza con la empresa Vitality, la aseguradora entrega smartbands de forma gratuita a sus asegurados. La entidad incentiva a sus asegurados a mantener un estilo de vida saludable, disminuyendo las reclamaciones. Además, ha generado alianzas con empresas de entretenimiento, viajes y compras, para obtener descuentos al cumplir con diferentes metas de ejercicio planteadas.

- Beam Digital

Beam Digital es una aseguradora dental que entrega cepillos “Smart”, los cuales monitorean la salud oral de sus clientes y envían reportes de sus hábitos de cepillado, esperando motivarlos a mejorar. La aseguradora aprovecha esta información para mejorar sus planes de seguros dentales y así reducir la prima cancelada por el seguro dental.

Los anteriores ejemplos son apenas una muestra del potencial que tiene el monitoreo de diferentes dispositivos de IoT en el sector asegurador tanto para la prevención de siniestros como para la fijación de un precio más competitivo lo que, potencialmente, permitiría a las aseguradoras ganar mayor presencia en el mercado. A su vez, vuelve el mercado y los precios de los productos más justos con aquellos individuos que adquieren los diferentes productos, ya que la fijación del precio se puede realizar en función de las características propias del tomador de la póliza y no un precio producto para un conjunto de características socioeconómicas y demográficas similares con, probablemente, un mayor nivel de riesgo del que realmente puede llegar a tener dicho cliente.

En Colombia, (SUMA movil, 2023) se estima que para 2025 el 21,09% de los hogares contarán con aparatos inteligentes. A pesar de ello, el sector asegurador recién ha iniciado a aprovechar esta enorme cantidad de datos y/o dispositivos. Por ejemplo:

- Seguros Falabella: en conjunto con SURA generó una póliza para autos denominada “Seguro Auto X km” en la cual se instala un dispositivo en el auto

que monitorea los kilómetros recorridos y, con esta información, el cobro se realiza en proporción a los mismos (Falabella, s.f.)

- Seguros Bolívar: Diseñó la póliza “Seguro de Autos por Recorridos”, que, de forma similar al anterior ejemplo, fue pensada para personas que utilizan poco su vehículo y desean cancelar solo el valor proporcional a lo recorrido. En este caso el tomador de la póliza puede activar desde su smartphone el inicio de la cobertura y desactivarla en el momento que llegue a su destino. Además, a través de un dispositivo Bluetooth se puede monitorear los hábitos de conducción que, por medio de diferentes alianzas, permiten acceder a descuentos u otro tipo de beneficios por el cumplimiento de metas (Seguros Bolivar, s.f.).

Basado lo anterior, existen grandes posibilidades de aplicación del IoT en el sector asegurador en Colombia en las líneas relacionadas a continuación, con un impacto considerable en la fijación de precios de las pólizas. Esto permitiría a las aseguradoras ganar una mayor participación en el mercado basado en pólizas más económicas que tengan en cuenta el riesgo real asumido por la aseguradora.

- Autos
  - Pay as you drive (Pago por km recorrido)
  - Pay How You Drive (por medio de sensores en el auto para conocer los hábitos de manejo del asegurado)
- Hogares
  - Sensores de humo
  - Drones en zonas de catástrofes
  - Cámaras digitales en el inmueble
  - Sensores de movimiento
  - Cerraduras digitales
  - Dispositivos de monitoreo de fugas de agua
- Salud
  - Monitoreo de la salud mediante Smartbands y Smartwatches
  - Cepillos de dientes inteligentes
- Ingeniería
  - Telemática en maquinarias
- Enfoque en el cliente
  - Chatbots
  - Pólizas “personalizadas”.

# 3. Problema

## 3.1. Planteamiento del problema

Tradicionalmente, los modelos de tarificación de pólizas de autos se basan en factores habituales, como la edad, modelo del carro, genero, ciudad de circulación, entre otras (FASECOLDA, s.f.), que puede llevar a una falta de precisión en la determinación del riesgo individual de cada cliente. Esta falta de precisión puede derivar en tarifas injustas o poco competitivas para ciertos clientes e incluso incrementar el riesgo de selección adversa, sin mencionar la falta de innovación en este tipo de productos por parte de las compañías aseguradoras.

Para abordar este problema, actualmente existen dispositivos IoT que pueden recopilar datos específicos del estado del vehículo, así como los hábitos de conducción de su usuario como aceleraciones o frenadas súbitas, sobre-revoluciones o cambios de marchas, etc. Si bien la implementación de este tipo de dispositivos plantea desafíos relacionados con la recopilación, análisis y gestión de grandes volúmenes de datos, así como preocupaciones sobre la privacidad y la seguridad de la información recopilada, esta data supone un insumo valioso para la generación de nuevos tipos de modelos de tarifación que permitan generar tarifas más justas para los clientes de acuerdo con su perfil de riesgo y/o hábitos de conducción.

En este sentido, se plantea como pregunta de investigación si incluir variables telemáticas (recopiladas por dispositivos IoT) en los modelos de tarifación de pólizas de autos, genera una prima pura de riesgo más acorde a los perfiles de riesgo de los asegurados.

Para responder a esta pregunta, se describe en este documento la metodología de adquisición de los datos, tanto reales como simulados, la generación de la base única de trabajo, la construcción de un modelo base de comparación y la generación de diferentes modelos que tienen en cuenta variables telemáticas suministradas por los dispositivos IoT. Finalmente se desarrolla las conclusiones del trabajo, y recomendaciones generales para la mejora continua de los resultados encontrados y, de esta forma, generar modelos más justos de tarifación que redunden, probablemente, en costos de pólizas más accesibles.

## 3.2. Árbol de problemas

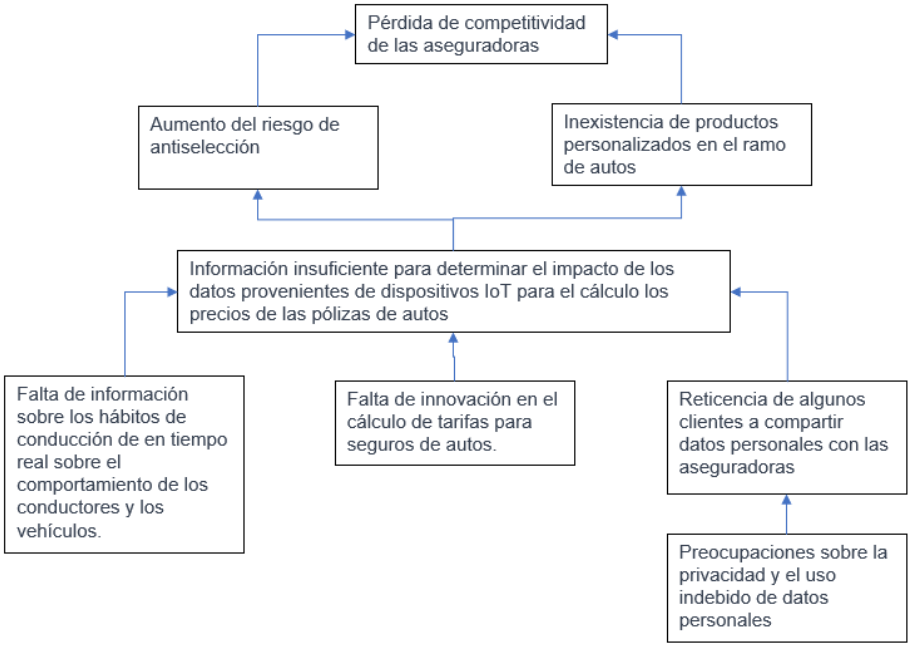


*Figura 1. Árbol de problemas*

## 3.3. Árbol de objetivos

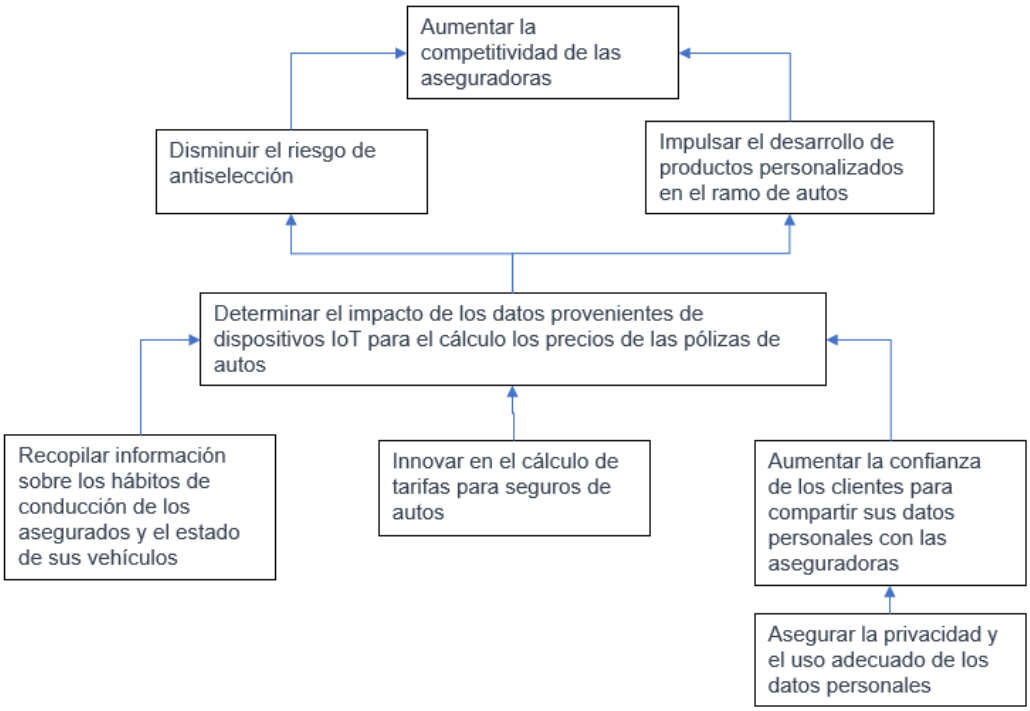


*Figura 2. Árbol de objetivos*

## 3.4. Objetivos

### 3.4.1. Objetivo general

Determinar el impacto del uso de datos provenientes de dispositivos IoT en el cálculo del precio de una póliza para el ramo de autos en una aseguradora de Colombia.

### 3.4.2. Objetivos específicos

- Recopilar información sobre los hábitos de conducción de asegurados con pólizas de autos
- Construir bases de datos con variables de tarifación tradicionales y extraídas de dispositivos IoT con comportamiento siniestral.
- Construir modelos de tarifación con variables tradicionales y con variables extraídas de dispositivos IoT.
- Comparar el ajuste de los modelos de tarificación construidos a los datos, para evaluar el impacto del uso de las variables telemáticas.

## 3.5. Alcances y limitaciones

### 3.5.1. Alcance

- El trabajo se enfoca en pólizas de autos individuales expedidas en Colombia.
- Se construye un modelo de tarificación utilizando variables tradicionales como punto de referencia para los demás modelos construidos.
- Los modelos de tarificación se construyen utilizando los modelos GLM (Generalized Linear Models).
- La comparación de los modelos construidos se realiza a través del estadístico AIC (Akaike Information Criterion), que es una medida utilizada para seleccionar el mejor modelo predictivo entre un conjunto de modelos candidatos.
- Se modela únicamente la frecuencia, severidad y tasa pura de riesgo, sin incluir, en ningún caso, el cálculo de tasas comerciales, ya que estas consideran factores independientes y propios de cada aseguradora.

### 3.5.2. Limitaciones

- La existencia de datos reales de telemática de asegurados en Colombia es una limitante, ya que la disponibilidad y calidad de los datos pueden afectar la construcción y precisión del modelo de tarificación.
- El número de registros contenidos en la base de datos unificada que contine variables tanto tradicionales como telemáticas puede resultar insuficiente para la construcción de los modelos, debido a las diferencias en las características de las poblaciones que representan.
- No se pretende utilizar metodologías alternas de modelación como RF (Random Forest), XGBoost o redes neuronales, lo que puede limitar la capacidad de los modelos para construidos para capturar patrones y relaciones más complejas en los datos.

## 4. Metodología

### 4.1. Definición de la metodología de trabajo

Para el desarrollo del trabajo se estableció la siguiente metodología:

- Adquisición de los datos: Búsqueda y adquisición de bases de datos reales o simuladas de pólizas de autos con variables tradicionales y telemáticas, es decir, obtenidas mediante dispositivos IoT.
- Tratamiento de las bases de pólizas tradicionales: Revisión y análisis inicial de las bases de pólizas tradicionales para definir rangos de las variables, variables a descartar, llaves entre las bases de pólizas y siniestros, etc.
- Unión de las bases tradicionales y telemáticas: Unión de la base de datos unificada de pólizas con variables tradicionales (pólizas + siniestros), con la base de pólizas con datos telemáticos.
- Alineación de la base construida con la base inicial: Selección aleatoria de registros en la base resultante de la unión de las bases de pólizas tradicionales y telemáticas con el fin de obtener una base final de trabajo y que su comportamiento sea acorde a la realidad del producto.
- Selección de variables: Definición de variables a incluir en los modelos a desarrollar que permitan eliminar problemas de correlación, multicolinealidad, dimensionalidad, etc.
- Segmentación de variables: Construcción de variables categóricas a partir de variables continuas, de acuerdo con las necesidades de los modelos definidos.
- Construcción de modelos: Construcción de por lo menos dos modelos de prima pura: uno específicamente con variables tradicionales y otro con variables tradicionales y variables telemáticas.
- Comparación de los modelos: Comparar los resultados de los modelos construidos para definir el de mejor ajuste a los datos. La medición del ajuste de

los modelos se realizará mediante la comparación de la deviance residual, el AIC (Akaike Information Criterion) o el BIC (Bayesian Information Criterion).

El desarrollo de cada una de las etapas mencionadas se realiza con mayor detalle a continuación.

## 4.2. Adquisición de datos

Para el desarrollo del presente trabajo de grado se requiere contar con bases de pólizas reales o simuladas que contengan la mayor cantidad de variables tradicionales para la construcción de modelos de pricing tales como: género, ciudad, edad del asegurado, modelo del vehículo, etc. Así mismo, y aún más importante para el desarrollo del trabajo, se requiere contar con bases de pólizas reales o simuladas que contengan la mayor cantidad de variables telemáticas obtenidas a través de dispositivos IoT, tales como: número de km recorridos, número de veces utilizado el vehículo, número de días utilizado a la semana, aceleraciones repentinas, frenadas repentinas, etc.

La base de datos de pólizas tradicionales adquirida consta de: un reporte de pólizas de autos reales expedidas en Colombia en el periodo 2020–01 a 2022–03 y otro reporte con las reclamaciones derivadas de estas mismas pólizas. Las bases contienen los siguientes campos y descripciones:

*Tabla 1. Campos de la base de pólizas reales*

| Campo | Descripción |
|---|---|
| **CODIGO COMPANIA** | Identificador único de la compañía de seguros que expidió la póliza. Campo anonimizado |
| **NUMERO POLIZA** | Identificador único de la póliza expedida. Campo anonimizado. |
| **FECHA INICIO VIGENCIA** | Fecha de inicio de vigencia de la póliza |
| **LUGAR RADICACION POLIZA** | Identificador del lugar donde se expidió la póliza. |
| **MODELO** | Modelo del auto asegurado |
| **PLACA** | Placa del vehículo asegurado |
| **CODIGO CIUDAD** | Identificador de la ciudad de expedición de la póliza |
| **TIPO DOCUMENTO ASEGURADO** | Identificador del tipo de documento del asegurado |
| **IDENTIFICACION ASEGURADO** | Número de documento del asegurado. Campo anonimizado |
| **SEXO** | Genero del asegurado |
| **PROFESION** | Código de la profesión del asegurado |
| **ESTADO CIVIL** | Estado civil del asegurado |
| **EDAD** | Edad del asegurado |
| **VALOR PRIMA PAGADA** | Valor de la prima pagada |
| **VALOR PRIMA SUSCRITA** | Valor de la prima suscrita |
| **VALOR ASEGURADO SIN RC** | Valor asegurado sin tener en cuenta la cobertura de responsabilidad civil |

| | |
|---|---|
| **PTD** | Marca de si la póliza cubre Pérdida Total Daños |
| **PPD** | Marca de si la póliza cubre Pérdida Parcial Daños |
| **PTH** | Marca de si la póliza cubre Pérdida Total Hurto |
| **PPH** | Marca de si la póliza cubre Pérdida Parcial Hurto |
| **RC** | Marca de si la póliza cubre Responsabilidad Civil |
| **VALOR ASEGURADO RC** | Valor asegurado teniendo en cuenta la cobertura de responsabilidad civil |
| **FECHA FIN VIGENCIA** | Fecha de finalización de vigencia de la póliza |
| **FECHA CANCELACION** | Fecha de cancelación de la póliza. Si aplica. |
| **FECHA MODIFICACION** | Fecha de última modificación de la póliza |

*Fuente: Base de datos de pólizas reales*

*Tabla 2. Campos de la base de siniestros reales*

| Campo | Descripción |
|---|---|
| **CODIGO COMPANIA** | Identificador único de la compañía de seguros que expidió la póliza. Campo anonimizado |
| **NUMERO SINIESTRO** | Identificador único del siniestro. Campo anonimizado. |
| **FECHA SINIESTRO** | Fecha en la que se presentó el siniestro |
| **CODIGO AMPARO** | Código del amparo/cobertura afectada |
| **VALOR RECLAMADO** | Valor del siniestro |
| **PLACA** | Placa del vehículo asegurado. Campo anonimizado |
| **SEXO** | Genero del asegurado |
| **EDAD** | Edad del asegurado |
| **PROFESION** | Código de la profesión del asegurado |
| **ESTADO CIVIL** | Estado civil del asegurado |
| **CODIGO CIUDAD** | Identificador de la ciudad de expedición de la póliza |
| **VALOR PRIMA PAGADA** | Valor de la prima pagada |
| **VALOR PRIMA SUSCRITA** | Valor de la prima suscrita |
| **VALOR ASEGURADO SIN RC** | Valor asegurado sin tener en cuenta la cobertura de responsabilidad civil |
| **MODELO** | Modelo del vehículo |
| **CIUDAD SINIESTRO** | Código de la ciudad donde se presentó el siniestro |
| **FECHA AVISO** | Fecha de aviso del siniestro |
| **FECHA PAGO** | Fecha de pago del siniestro |
| **VALOR RESERVA CONSTITUIDA** | Valor de la reserva constituida |
| **VALOR RESERVA PAGADA** | Valor de la reserva pagada |
| **VALOR PAGADO** | Valor pagado |
| **VALOR SINIESTRO INCURRIDO** | Valor incurrido del siniestro |

| | |
|---|---|
| **NUMERO SINIESTROS ANTERIORES** | Número de siniestros presentados anteriores a la presente reclamación |
| **CODIGO DEPARTAMENTO** | Código del departamento donde se presentó el siniestro |
| **COLOR** | Color del vehículo |
| **VALOR ASEGURADO RC** | Valor asegurado teniendo en cuenta la cobertura de responsabilidad civil |
| **NUMERO POLIZA** | Identificador único de la póliza expedida. Campo anonimizado. |

*Fuente: Base de siniestros reales*

Debido a la novedad del tema y la confidencialidad de la información, no fue posible adquirir bases de datos reales de pólizas con variables telemáticas en Colombia. Por lo tanto, se utilizó una base de datos simulados de pólizas que permitiese llevar a cabo el objetivo. Se utilizó la base de datos de la publicación "Synthetic Dataset Generation of Driver Telematics" (So, Boucher, & Valdez, 2021), cuyo resultado principal consistió en la construcción de una base sintética de datos de pólizas de autos con variables tradicionales y telemáticas de asegurados en Canadá, con el fin de utilizarla para diferentes investigaciones sobre modelación de riesgos. Las variables contenidas en el mencionado artículo corresponden a:

*Tabla 3. Campos incluidas en la base de datos simulada*

| Tipo de variable | Campo | Descripción |
|---|---|---|
| **Tradicional** | Duration | Duración en días de la cobertura de la póliza |
| | Insured.age | Edad del asegurado, en años |
| | Insured.sex | Género del asegurado |
| | Car.age | Número de años del vehículo |
| | Car.use | Tipo de uso que se le da al vehículo: Privado, Trabajo, Agrícola, Comercial |
| | Credit.score | Puntaje de crédito del asegurado |
| | Region | Tipo de región donde se utiliza el vehículo: Urbano, Rural |
| | Annual.miles.drive | Millas anuales que se espera utilizar el vehículo, según lo declarado por el asegurado |
| | Years.noclaims | Número de años sin ningún reclamo |
| | Territory | Localización territorial del vehículo |
| **Telemáticas** | Annual.pct.driven | Porcentaje anualizado de tiempo en uso |
| | Total.miles.driven | Distancia total recorrida, en años |
| | Pct.drive.xxx | Porcentaje de uso el día XXX de la semana |
| | Pct.drive.xhrs | Porcentaje de uso del vehículo de menos de X horas |

| | | |
|---|---|---|
| | Pct.drive.xxx | Porcentaje de manejo del vehículo: Entre semana o fines de semana |
| | Pct.drive.rushXXX | Porcentaje de manejo durante XXX horas pico: am/pm |
| | Avgdays.week | Número promedio de días utilizado por semana |
| | Accel.XXmiles | Número de aceleraciones súbitas 6/8/9/…/14 mph/s por cada 1000 millas |
| | Brake.XXmiles | Número de frenadas súbidas 6/8/9/…/14 mph/s por cada 1000 millas |
| | Left.turn.intensityXX | Número de giros a la izquierda por cada 1000 millas con intensidad de 08/09/10/11/12 |
| | Right.turn.intensityXX | Número de giros a la derecha por cada 1000 millas con intensidad de 08/09/10/11/12 |
| **Respuesta** | NB_Claim | Número de reclamos durante el periodo de observación |
| | AMT_Claim | Valor agregado de los reclamos durante el periodo de observación |

*Fuente: (So, Boucher, & Valdez, 2021)*

En adelante, el documento se refiere a las bases de pólizas reales y sus siniestros como “bases reales” y a la base de información telemática como “base simulada”.

## 4.3. Tratamiento de las bases reales

Como se mencionó anteriormente, las bases de datos de pólizas reales se dividen en 2: La información de las pólizas y la información de los siniestros.

La base de información de las pólizas consta de 1.401.972 registros y 44 campos de los cuales se seleccionó: fecha inicio vigencia, fecha fin vigencia, edad, sexo, modelo, estado civil, código compañía, numero póliza, placa, valor prima suscrita y valor asegurado sin RC, con el fin de alinear esta base con la base de siniestros y la base simulada. Al tomar valores únicos de las pólizas resultantes, se obtiene una base con 1.400.403 registros.

La base de información de los siniestros consta de 217.900 registros y 64 campos de los cuales se seleccionó: código compañía, placa, numero póliza, numero siniestro y valor pagado, con el fin de alinear esta información con la base de la información de las pólizas para generar una base de datos reales unificados. A su vez, se agruparon los resultados mediante la suma de los diferentes identificadores de siniestros y el valor pagado por los mismos. Finalmente, realizó un filtro para eliminar aquellos registros que contaran con reclamos cuyo valor pagado sea 0. Se obtiene una base de 44.080 registros.

La unión de las bases de pólizas y siniestros se realizó a través de la llave código compañía, numero póliza y placa, obteniendo una base inicial unificada de 1.400.003 registros. Sobre esta base se realizó un análisis para determinar el periodo de desarrollo de los siniestros:

*Tabla 4. Factores de desarrollo de pólizas reales*

| Periodo de desarrollo | Factor |
|---|---|
| **De 0 a 1 meses** | 1,1619 |
| **De 1 a 2 meses** | 1,0372 |
| **De 2 a 3 meses** | 1,0126 |
| **De 3 a 4 meses** | 1,0112 |
| **De 4 a 5 meses** | 1,0047 |
| **De 5 a 6 meses** | 1,0032 |
| **De 6 a 7 meses** | 1,0053 |
| **De 7 a 8 meses** | 1,0038 |
| **De 8 a 9 meses** | 1,0033 |
| **De 9 a 10 meses** | 1,0003 |
| **De 10 a 11 meses** | 1,0001 |
| **De 11 a 12 meses** | 1,0011 |
| **De 12 a 13 meses** | 1,0019 |
| **De 13 a 14 meses** | 1,0049 |
| **De 14 a 15 meses** | 1,0001 |
| **De 15 a 16 meses** | 1,0000 |
| **De 16 a 17 meses** | 1,0000 |
| **De 17 a 18 meses** | 1,0008 |
| **De 18 a 19 meses** | 1,0000 |
| **De 19 a 20 meses** | 1,0015 |

*Fuente: Elaboración propia*

En la Tabla 4 se observa que la mayor parte de los siniestros se desarrolla casi en su totalidad en los primeros 12 meses después del aviso. Con el fin de incluir la mayor cantidad de datos posible, se tomó una ventana de pólizas expedidas entre 2020-01 y 2021-03.

## 4.4. Criterios de inclusión y selección de variables

Se generaron los siguientes filtros de modo que la base construida sea lo más limpia posible:

- Se incluyeron pólizas expedidas entre 2020-01 y 2021-03
- Se excluyeron pólizas con fecha de fin anterior a la fecha de inicio.
- Se excluyeron pólizas donde la edad es negativa o desconocida
- Se excluyeron pólizas donde el género es desconocido
- Se excluyeron pólizas donde el estado civil es desconocido
- Se excluyeron campos sin información
- Se excluyeron aquellos registros que contaran con reclamos cuyo “valor pagado” sea 0.
- Se excluyeron variables duplicadas o irrelevantes para la construcción del modelo (p.e. Identificadores únicos)

- Se excluyeron campos altamente correlacionados. Se mantiene el campo de mayor aporte al componente principal número 1.

Finalmente, se calculó el porcentaje de siniestros general de la base, obteniendo un total de 632.393 pólizas con 20.069 siniestros; es decir, una tasa de reclamos del 3,17%.

## 4.5. Unión de las bases real y simulada

Una vez construida la base real consolidada, se realizó la unión con la base simulada para obtener una base única de trabajo que permitiera el desarrollo de los modelos de tarifación.

Dado que no existe una llave natural entre la base real consolidada y la base simulada, se decidió unirlas a partir de los campos mencionados a continuación, partiendo del supuesto que, por ejemplo, una persona soltera, masculina, de 21 años con vehículo de 3 años de antigüedad y que ha tenido 2 siniestros, se comporta de forma similar tanto en los datos reales como en los datos simulados:

- Insured age: Edad del asegurado
- Insured sex: Género del asegurado
- Car age: Antigüedad del vehiculo
- Marital: Estado civil del asegurado
- NB: Claim: Número de reclamaciones / siniestros

Sin embargo, al realizar la unión de esta forma, se evidenciaron registros duplicados. Para solucionar esta situación se realizó el siguiente proceso:

1. Se agregó una columna ID (único) a las bases real consolidada y a la base simulada
2. Se unió la información de la base de datos simulada a la base de datos real mediante un left join
3. La base generada se organiza por ID de la base real y luego por ID de la base simulada. Se construye una nueva columna que enumere el número de repeticiones del ID de la base real.
4. Se filtra la base para tomar únicamente la primera repetición por ID de la base real y se eliminan las demás repeticiones. De esta forma se asegura que no se dupliquen los registros.
5. Se guardan los resultados en una base denominada “base unida”
6. Se generan bases pivote donde se eliminan los registros que ya han cruzado tanto en la base real como en la base simulada.
7. Se repiten los pasos 2 a 6 hasta que ya no se obtengan cruces entre las bases

Como resultado del proceso, de las 632.393 pólizas de la base real y de las 100.000 pólizas de la base simulada, se construyó una base unificada con 92.096 pólizas con variables

tradicionales y telemáticas que les corresponde 1.123 reclamos; es decir, una tasa del 1.22%.

## 4.6. Alineación de la base construida con la base real

La base unificada construida, difiere significativamente (en su tasa de reclamos) de la base real original. Por ello, se realizó una selección aleatoria de registros de modo que la base unificada obtenida se comporte, en general, como la base real. Así mismo, se conservaron todos los registros con reclamos, que son los principales datos que permitirán realizar la construcción de los modelos.

*Tabla 5.Conteo de registros por número de reclamos*

| Reclamos | Registros |
|---|---|
| **0** | 90.985 |
| **1** | 1.099 |
| **2** | 12 |
| **3** | 0 |
| **Total** | 92.096 |
| **Tasa de reclamos** | 1,22% |

La selección aleatoria se realizó sobre los registros con 0 reclamos a partir de una regla de 3 simple, es decir, se seleccionaron aleatoriamente 34.276 registros. La base final quedó configurada de la siguiente forma:

*Tabla 6. Conteo de registros por número de reclamos de la base unificada*

| Reclamos | Registros |
|---|---|
| **0** | 34.276 |
| **1** | 1.099 |
| **2** | 12 |
| **3** | 0 |
| **Total** | 35.387 |
| **Tasa de reclamos** | 3,17% |

De este modo se obtiene una base final unificada con 35.387 registros con una tasa de reclamos del 3,17%, igual que la base real original.

## 4.7. Selección de variables

La base unificada consta de 66 campos, a saber:

```
> names(Base_Unida_final)
 [1] "Duration.x"              "Insured.age"             "Insured.sex"             "Car.age"
 [5] "Marital"                 "Car.use.x"               "Credit.score.x"          "Region.x"
 [9] "Years.noclaims.x"        "Territory.x"             "NUM_RECLAMOS"            "VALOR_PAGADO_2"
[13] "Premium"                 "Exposure"                "llave"                   "NB_Claim"
[17] "AMT_Claim.x"             "id.x"                    "Duration.y"              "Car.use.y"
[21] "Credit.score.y"          "Region.y"                "Annual.miles.drive"      "Years.noclaims.y"
[25] "Territory.y"             "Annual.pct.driven"       "Total.miles.driven"      "Pct.drive.mon"
[29] "Pct.drive.tue"           "Pct.drive.wed"           "Pct.drive.thr"           "Pct.drive.fri"
[33] "Pct.drive.sat"           "Pct.drive.sun"           "Pct.drive.2hrs"          "Pct.drive.3hrs"
[37] "Pct.drive.4hrs"          "Pct.drive.wkday"         "Pct.drive.wkend"         "Pct.drive.rush am"
[41] "Pct.drive.rush pm"       "Avgdays.week"            "Accel.06miles"           "Accel.08miles"
[45] "Accel.09miles"           "Accel.11miles"           "Accel.12miles"           "Accel.14miles"
[49] "Brake.06miles"           "Brake.08miles"           "Brake.09miles"           "Brake.11miles"
[53] "Brake.12miles"           "Brake.14miles"           "Left.turn.intensity08"   "Left.turn.intensity09"
[57] "Left.turn.intensity10"   "Left.turn.intensity11"   "Left.turn.intensity12"   "Right.turn.intensity08"
[61] "Right.turn.intensity09"  "Right.turn.intensity10"  "Right.turn.intensity11"  "Right.turn.intensity12"
[65] "AMT_Claim.y"             "id.y"
```

*Figura 3. Campos de la base unificada*

Aquellas variables con sufijo ".x" o ".y" corresponden a campos de información que se encuentran tanto en la base real como en la base simulada, respectivamente.

Un filtro sobre esta base consistió en eliminar las variables duplicadas o irrelevantes para la construcción del modelo. Las variables eliminadas en esta etapa son:

| # | Variable eliminada | Razón |
|---|---|---|
| **1** | Tipo de uso | Campo sin información en la data real |
| **2** | Puntaje crediticio | Campo sin información en la data real |
| **3** | Region de uso | Campo sin información en la data real |
| **4** | Años sin reclamos | Campo sin información en la data real |
| **5** | Territorio | Campo sin información en la data real |
| **6** | Numero de reclamos | Campo duplicado. Se mantiene NB_Claim |
| **7** | Valor pagado | Campo duplicado. Se mantiene AMT_Claim |
| **8** | Llave | Identificador para realizar la unión de las bases de pólizas reales con la de sus reclamos. |
| **9** | Id | Identificador único utilizado para la unión de las bases real y simulada |
| **10** | Valor de los reclamos (data simulada) | Corresponde al valor pagado por los reclamos de la base simulada. Se mantiene la variable AMT_Claim.x que corresponde al valor pagado por los reclamos en la base real |

Se analizaron las variables telemáticas, con el fin de eliminar variables altamente correlacionadas que puedan generar problemas de multicolinealidad en la construcción del modelo. El siguiente gráfico muestra el nivel de correlación presentado entre cada una de las variables (Figura 4).

Se revelaron altas correlaciones entre diferentes conjuntos de variables, como por ejemplo entre "Annual.pct.driven" y "Total.miles.driven" o entre las variables que indican el uso del

carro los días lunes, martes, miércoles, jueves y viernes con la variable que indica el uso del carro entre semana.

Se observó también agrupaciones naturales de variables, por lo que se realizó un análisis de los siguientes grupos: “pct.drive.”, “Accel.”, “Brake.”, “Right.” y “Left.”.

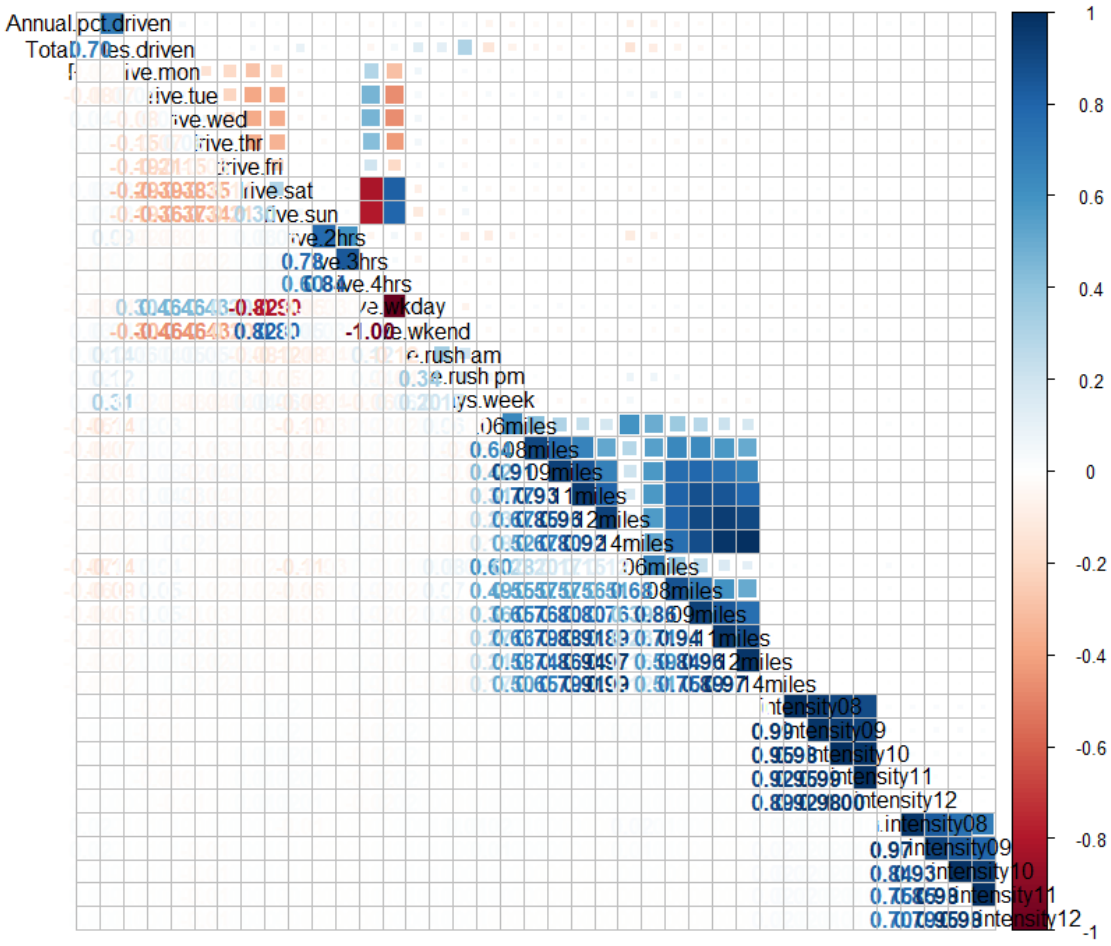


*Figura 4. Correlación entre las variables telemáticas. Elaboración propia*

A continuación, se presentan las variables con una correlación mayor o igual, en términos absolutos, al 0.7, para facilidad de interpretación:

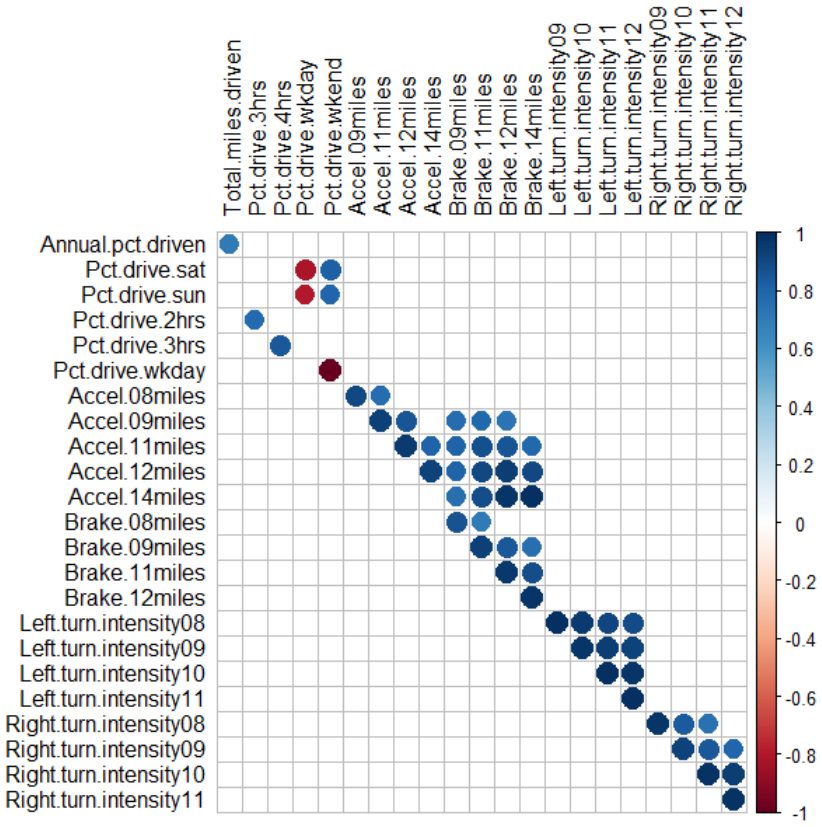


*Figura 5. Variables telemáticas con correlación superior a 0.7. Elaboración propia*

Para la selección de las variables más importantes por cada grupo de variables, se analizó la correlación entre las mismas. Posteriormente, se realizó la construcción de los componentes principales y escogió aquellas variables que más aportan en el componente principal 1. El propósito de esta etapa fue escoger entre 1 y 2 variables relevantes por grupo, para la construcción de los modelos, de manera que se eliminaran posibles problemas de multicolinealidad.

- Grupo 1: Variables Pct.drive.

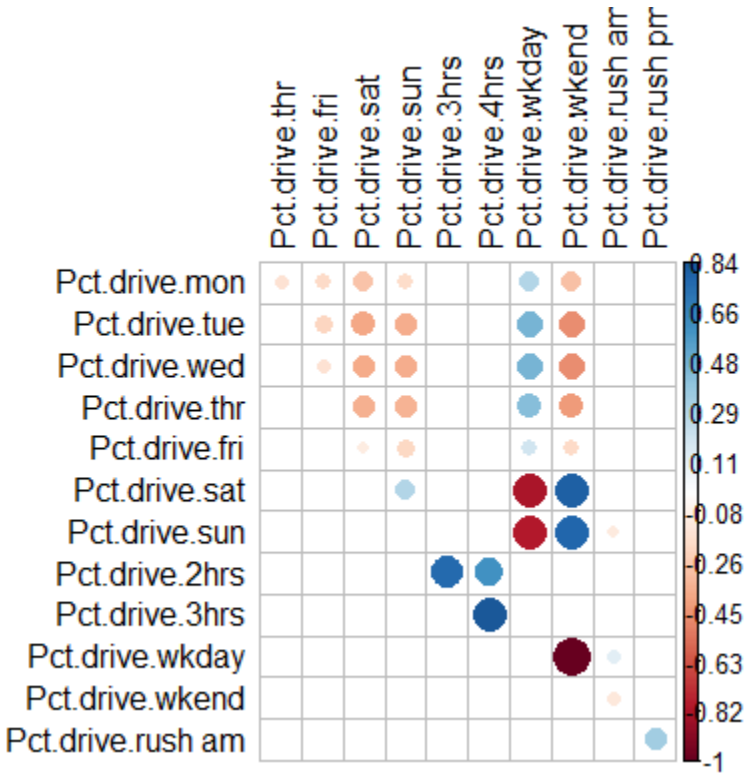


*Figura 6. Correlación de las variables Pct.drive. Elaboración propia*

Como se presentó anteriormente, se observan fuertes correlaciones entre algunas de las variables del grupo. Para decidir las variables a utilizar en el modelo y las variables a descartar se realizó un análisis de componentes principales. De acuerdo con (Williams & Abdi, 2010), el análisis de componentes principales (PCA por sus siglas en inglés) es una técnica multivariada de análisis de datos que se utiliza principalmente para reducir la dimensionalidad de conjuntos de datos. La idea principal detrás de PCA es encontrar las direcciones de mayor variabilidad en los datos, conocidas como los componentes principales, y proyectar los datos originales en estas nuevas dimensiones.

En particular, cuando se tiene un conjunto de datos con muchas variables es posible que algunas de ellas sean redundantes o irrelevantes, lo que puede dificultar el análisis y la interpretación de los resultados. En este caso, el PCA permite reducir el número de variables mientras se mantiene la mayor cantidad de información posible, lo que facilita la visualización y la comprensión de los patrones subyacentes en los datos, así como eliminar la multicolinealidad que pueda aparecer entre los mismos.

Como resultado de aplicar PCA al conjunto de datos se observa que las variables “pct.drive.wkday” y “pct.drive.wkend” son las de mayor contribución al componente principal 1, por consiguiente, se descartan las demás variables del grupo.

Adicionalmente, se eliminó la variable “pct.drive.wkend” debido a que corresponde al complemento de la variable “pct.drive.wkday”, lo que genera problemas de colinealidad en el momento de la construcción de los modelos.

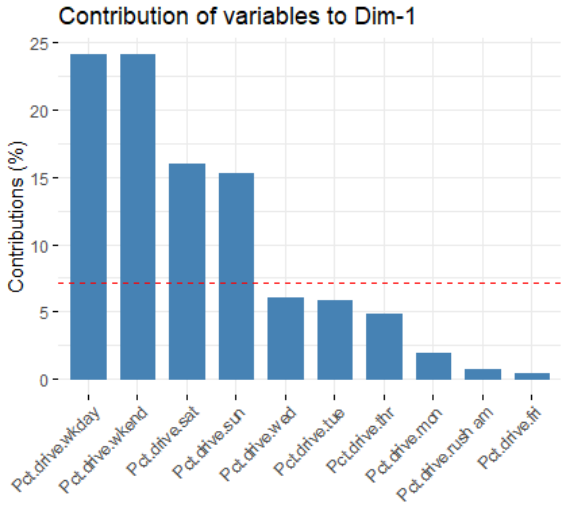


*Figura 7. Contribución de las variables Pct.drive a la dimensión 1 del PCA*

- Grupo 2: Variables Accel.

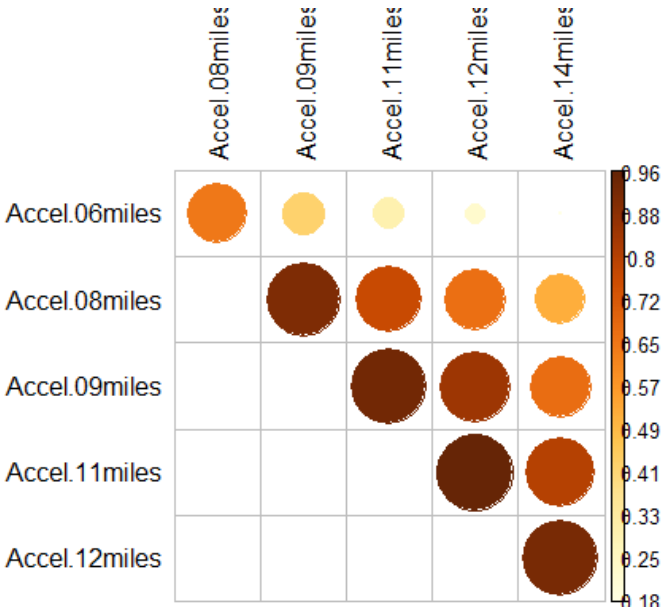


*Figura 8. Correlación de las variables Accel. Elaboración propia*

Para la generación del PCA no se requirió realizar ningún pre-procesamiento de las variables del grupo, dado que se encuentran en la misma unidad de medida (Shlens, 2003).

El resultado del PCA indica que las variables “Accel.11.miles” y “Accel.09.miles” son las de mayor contribución al componente principal 1, por consiguiente, se descartan las demás variables del grupo.

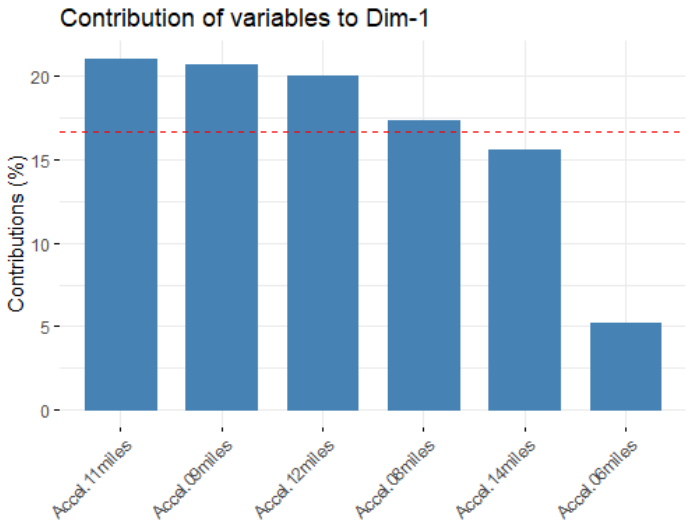


*Figura 9. Contribución de las variables Accel a la dimensión 1 del PCA*

- Grupo 3: Variables Brake.

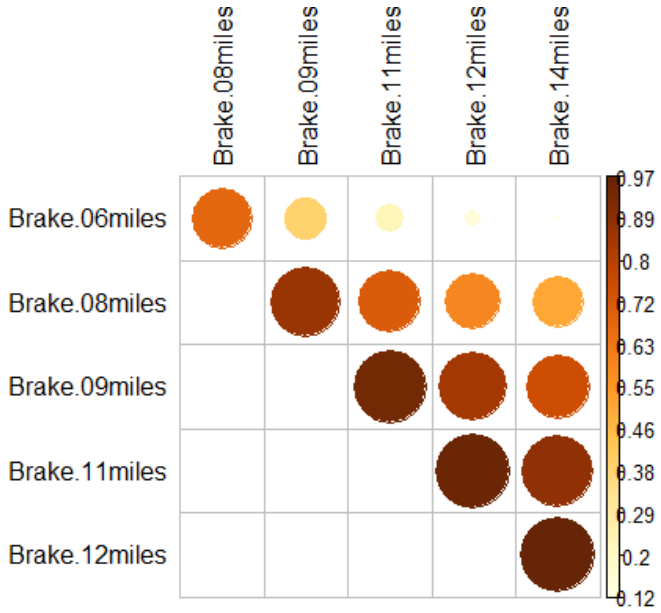


*Figura 10. Correlación de las variables Brake. Elaboración propia*

El resultado del PCA indica que las variables “Brake.11.miles” y “Brake.09.miles” son las de mayor contribución al componente principal 1, por consiguiente, se descartan las demás variables del grupo.

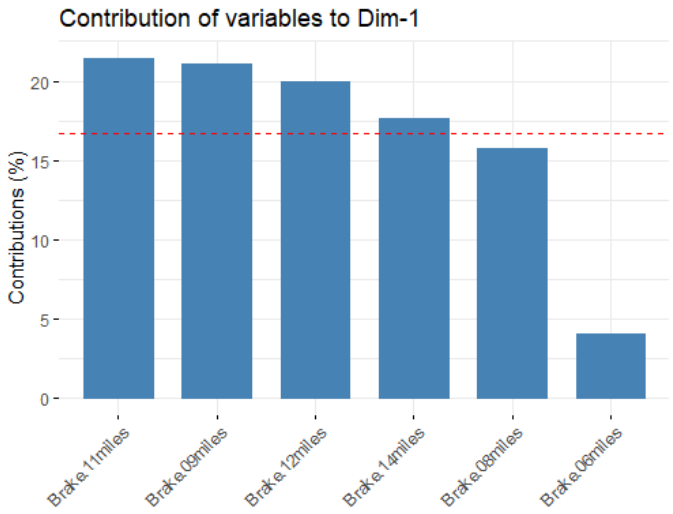


*Figura 11. Contribución de las variables Brake a la dimensión 1 del PCA. Elaboración propia*

- Grupo 4: Variables Right.

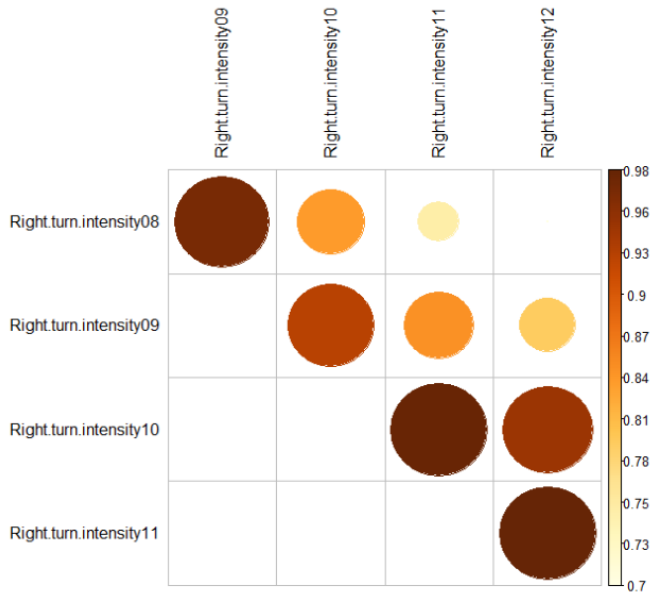


*Figura 12. Correlación de las variables Right. Elaboración propia*

El resultado del PCA indica que la variable “Right.turn.intensity.10” es la de mayor contribución al componente principal 1, por consiguiente, se descartan las demás variables del grupo.

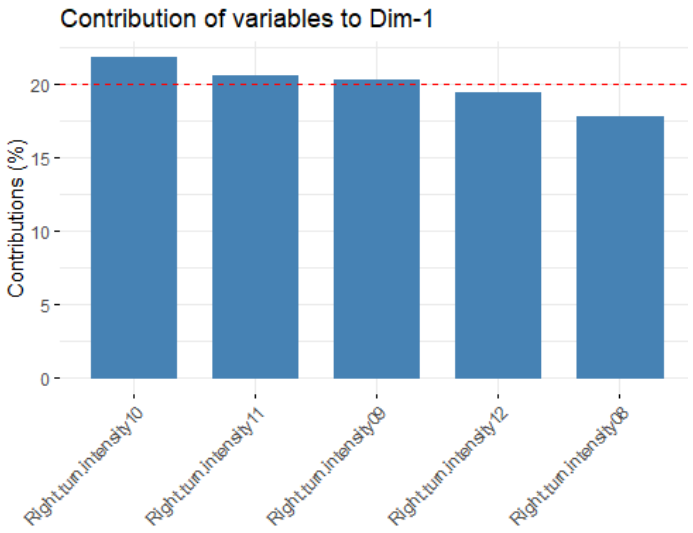


*Figura 13. Contribución de las variables Right a la dimensión 1 del PCA. Elaboración propia*

- Grupo 5: Variables Left.

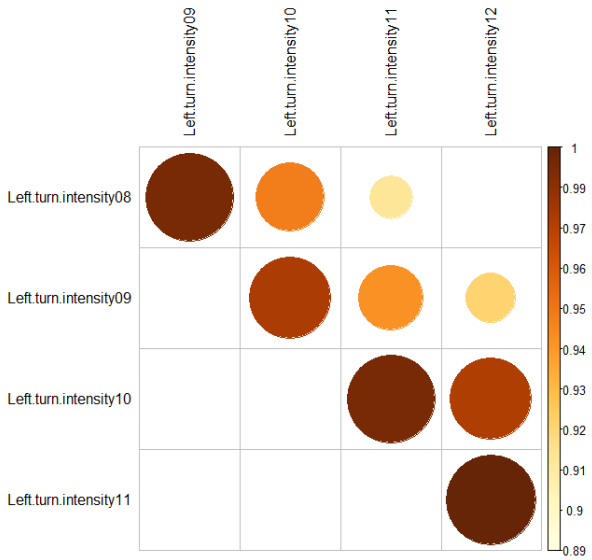


*Figura 14. Correlación de las variables Left. Elaboración propia*

Para la generación del PCA no se requiere realizar ningún pre-procesamiento de las variables del grupo, dado que se encuentran en la misma unidad de medida.

El resultado del PCA se muestra a continuación en donde se observa que la variable "LeftTurnIntensity.10" es la de mayor contribución al componente principal 1, por consiguiente, se descartan las demás variables del grupo.

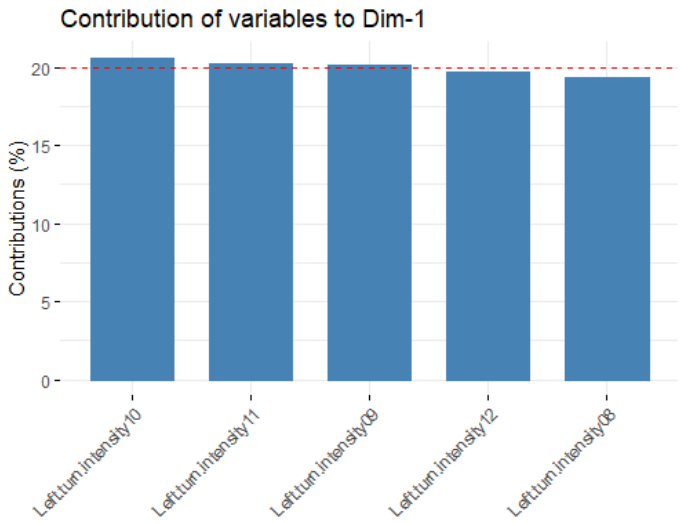


*Figura 15. Contribución de las variables Left a la dimensión 1 del PCA. Elaboración propia*

Como conclusión, a través de un análisis de correlación y con el apoyo de los componentes principales, se seleccionan las siguientes variables, de modo que la pérdida de información sea mínima, se eliminen posibles problemas de colinealidad y de paso se reduzca la dimensionalidad de la data, para que no se vean penalizados los modelos construidos.

## 4.8. Segmentación de variables

Se realiza la segmentación de las siguientes variables, con el fin de prepararlas para la construcción de los modelos:

- Insured.age: De acuerdo con el (Ministerio de salud y protección social, 2023) "El ciclo vital puede dividirse en diferentes etapas del desarrollo (…). La siguiente clasificación es un ejemplo: (…) juventud (14 - 26 años), adultez (27 - 59 años) y vejez (60 años y más)."
  El histograma de la frecuencia de edades, de acuerdo con los grupos mencionados anteriormente permite apreciar una mayor tasa de reclamaciones en la población

joven, luego un descenso controlado de la tasa de reclamaciones en la población adulta y finalmente un descenso más pronunciado en la población “vieja”.

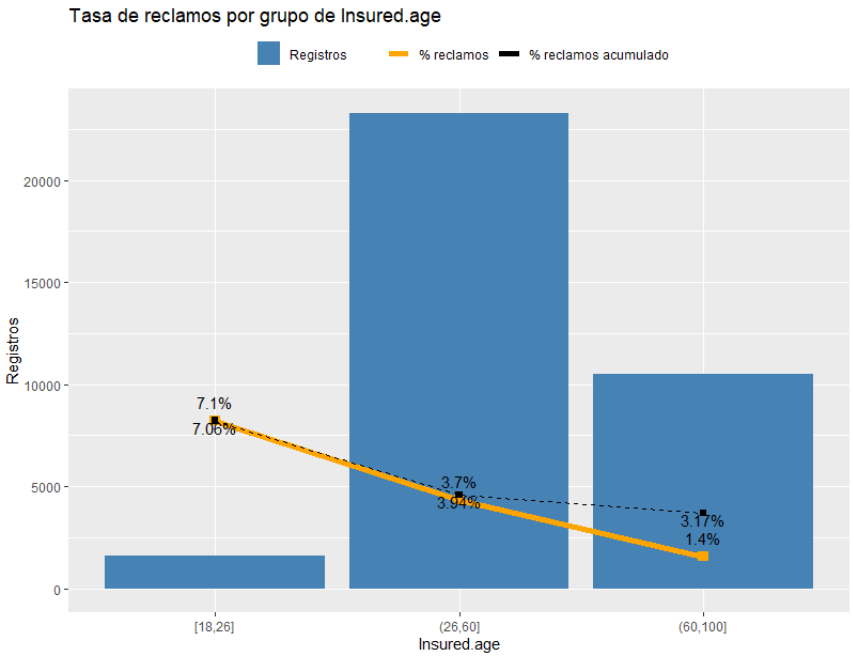


*Figura 16. Tasa de reclamos por categorías de Insured.age. Elaboración propia*

- Car.age: Desde hace varios años, las entidades financieras han otorgado plazos de hasta 5 años para el pago de créditos con destinación a la compra de vehículos nuevos o usados (Semana, 2023) y sobre los que se debe adquirir pólizas todo riesgo, al menos por el plazo del crédito, de aquí que la mayor proporción de registros, se encuentren en el rango -2 a 5. En este sentido, se generan las siguientes categorías para la variable: Vehículos de hasta 5 años de antigüedad, vehículos de 6 a 9 años de antigüedad y vehículos de 10 o más años de antigüedad. A continuación, se presenta el histograma de la frecuencia de antigüedad del vehículo, junto con la tasa de reclamos puntual y acumulada.

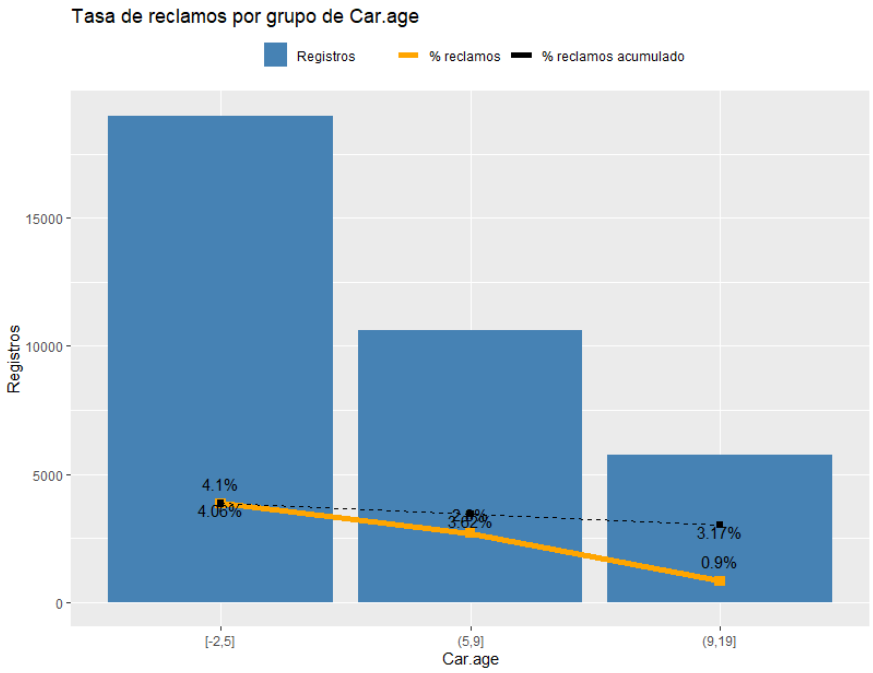


*Figura 17. Tasa de reclamos por categoría de Car.age. Elaboración propia*

- years.noclaims: Se realizó un análisis descriptivo de la variable para encontrar particiones que permitan representar proporciones representativas de la población y que la tasa de reclamos sea diferencial entre grupos. En este caso, se generan las categorías de 1 a 10, de 10 a 35, de 35 a 55 y de 55 a 75, donde cada categoría contiene un número representativo de registros y además la tasa de reclamos en cada grupo es sustancialmente diferencial de los demás.

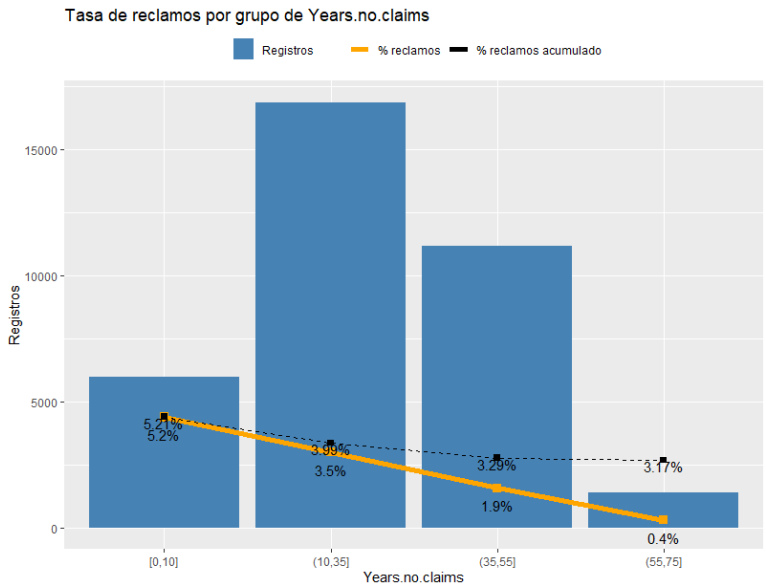


*Figura 18. Tasa de reclamos por categoría de Years.noclaims. Elaboración propia*

- pct.drive.wkday: Se construyó las categorías de 1 a 0.7, de 0.7 a 0.8 y de 0.8 a 1, donde cada categoría contiene un número representativo de registros y además la tasa de reclamos en cada grupo es sustancialmente diferencial de los demás.

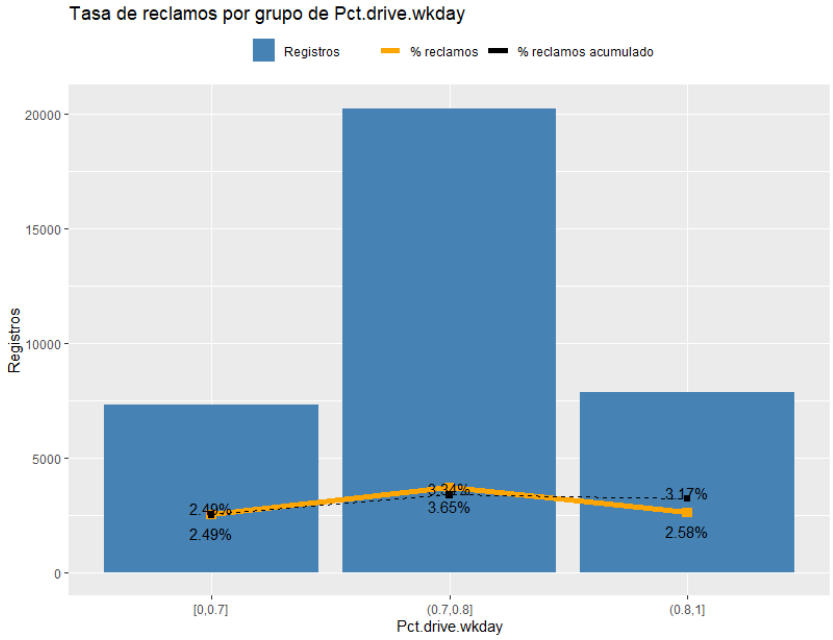


*Figura 19. Tasa de reclamos por categoría de Pct.drive.wkday. Elaboración propia*

- Accel.09miles: Se construyó las categorías: 0, de 1 a 5 y de 6 en adelante con el fin de conseguir grupos con tasas de reclamos diferenciales.

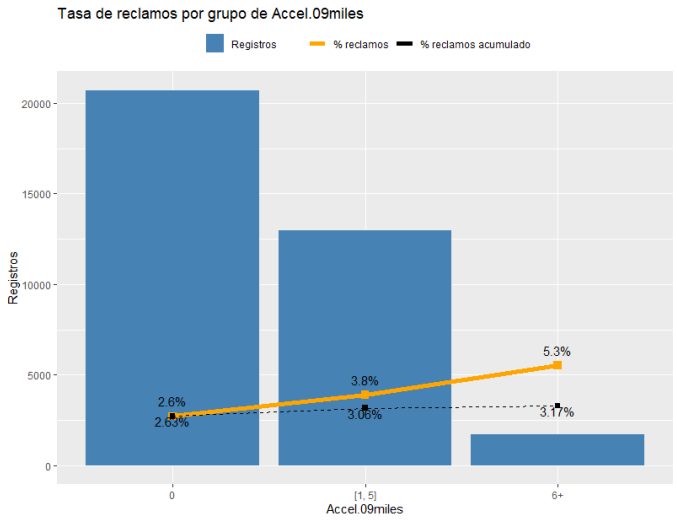


*Figura 20. Tasa de reclamos por categoría de Accel.09miles. Elaboración propia*

- Accel.11miles: Se construyó las categorías: 0, de 1 a 3 y de 4 en adelante, donde cada categoría indica una tasa de reclamos diferencial.

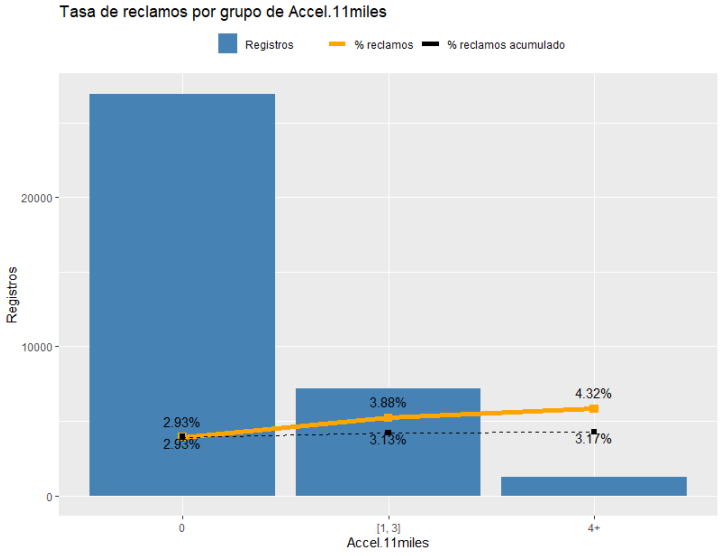


*Figura 21. Tasa de reclamos por categoría de Accel.11miles. Elaboración propia*

- Brake.09mile: Se construyó las categorías: 0, de 1 a 7 y de 8 en adelante, donde cada categoría contiene un número representativo de registros y además la tasa de reclamos en cada grupo es sustancialmente diferencial de los demás.

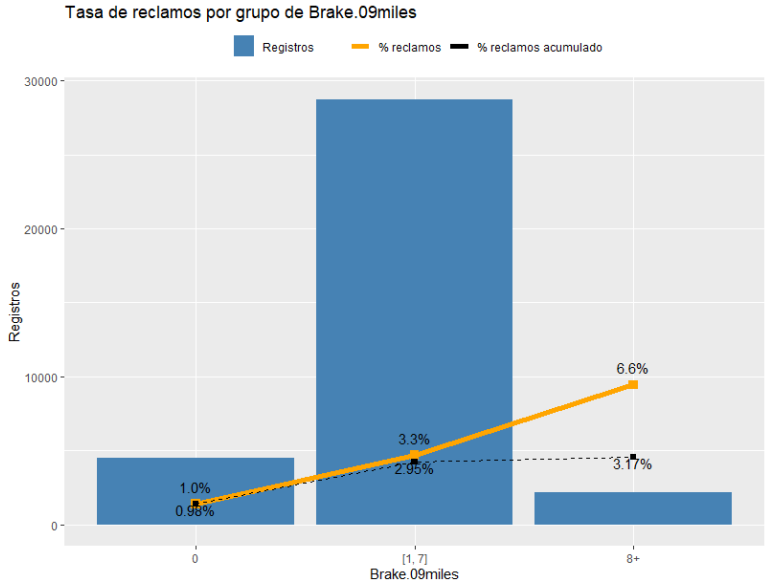


*Figura 22. Tasa de reclamos por categoría de Brake.09miles. Elaboración propia*

- Brake.11miles: Se construyó las categorías: 0, 1 y de 2 en adelante. Se observa un fuerte incremento en la tasa de reclamos en el grupo “de 2 en adelante”.

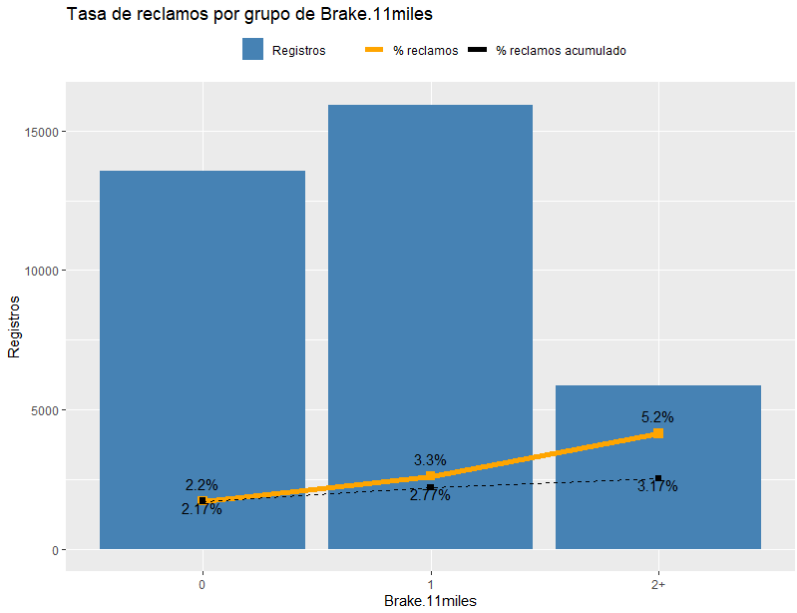


*Figura 23. Tasa de reclamos por categoría de Brake.11miles. Elaboración propia*

- Left.turn.intensity10: Se construyó las categorías: 0, de 1 a 10 y de 11 en adelante.

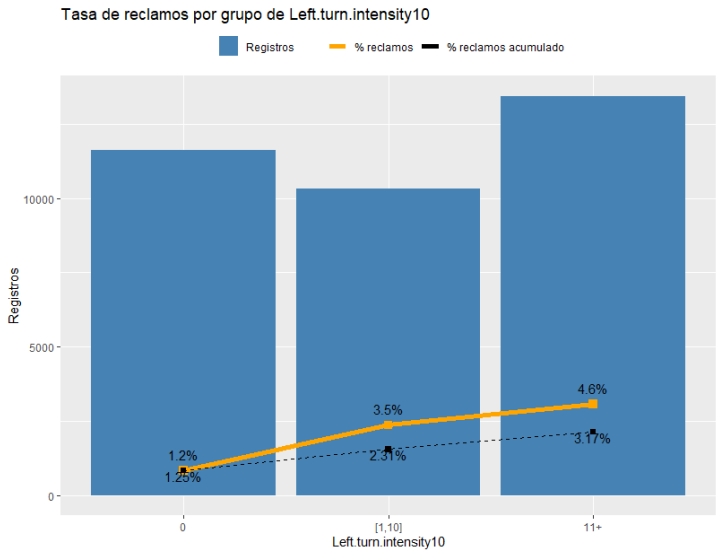


*Figura 24. Tasa de reclamos por categoría de Left.turn.intensity10. Elaboración propia*

- Right.turn.intensity10: Se construyó las categorías: 0, de 1 a 20 y de 21 en adelante. Se observa que la tasa de reclamos de la categoría "0" es notablemente inferior a las demás categorías.

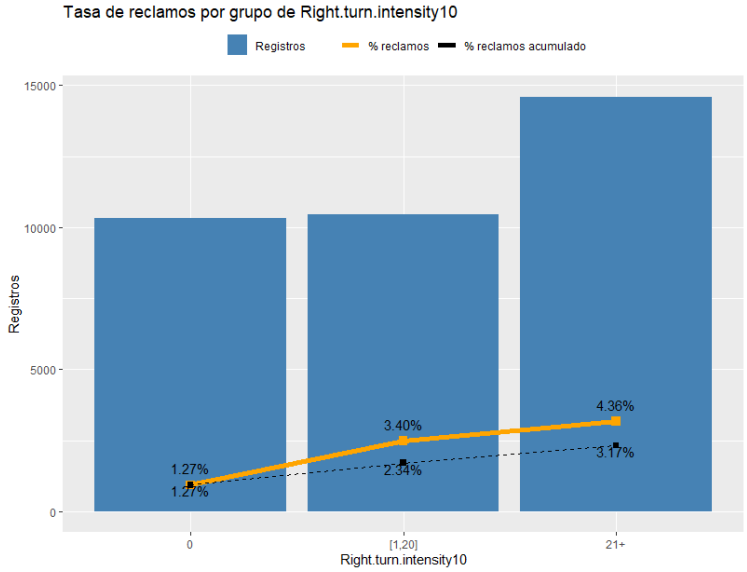


*Figura 25. Tasa de reclamos por categoría de Right.turn.intensity10. Elaboración propia*

## 4.9. Construcción de modelos

Una vez definidas las variables a tener en cuenta para la construcción de los modelos, se procedió a construir los modelos de prima pura tanto con las variables tradicionales como con las variables telemáticas obtenidas de los dispositivos IoT. Se omitió en esta parte el

cálculo de la prima comercial ya que este concepto varía de aseguradora a aseguradora, y no permite hacer una comparación real del ajuste de los modelos construidos.

La construcción de los modelos de prima pura se realizó utilizando Modelos Lineales Generalizados (GLM por sus siglas en inglés) que son un tipo de modelo estadístico usado para modelar una variable a predecir o *variable objetivo* que no necesariamente se encuentra distribuida normalmente (como en los modelos lineales) pero que debe pertenecer a la "familia de distribuciones exponenciales". (Anderson, 2023)

Así mismo, los modelos GLM son ampliamente utilizados en la industria aseguradora por su buen desempeño, capacidad de explicación de las variables, entre otras, cuando la variable objetivo suele ser una de las siguientes (Goldburd, Khare, Tevet, & Guller, 2020):

- Frecuencia de siniestros (es decir, siniestros por exposición)
- Gravedad de los siniestros (valor de la pérdida por siniestro o suceso)
- Prima pura (es decir, valor de la pérdida por exposición)
- Ratio de siniestralidad (valor de los siniestros por valor de la prima)

Para variables objetivo de naturaliza cuantitativas como las anteriores, los GLM's generan una estimación del valor esperado del resultado.

Ahora bien, la construcción de los modelos de tarifación se puede realizar, principalmente, de 2 formas:

1. Construir modelos de Frecuencia / Número de reclamos y Severidad / Valor de los reclamos:

   En este caso, los modelos individuales se combinan para formar un modelo de prima pura. Por ejemplo: Suponiendo que se hayan utilizado modelos GLM con enlaces logarítmicos para ambos, esta combinación de los dos modelos se consigue simplemente multiplicando sus correspondientes factores de relatividad (Goldburd, Khare, Tevet, & Guller, 2020).

La modelación de la frecuencia de los siniestros se realiza principalmente mediante dos distribuciones: la distribución de Poisson y la distribución binomial negativa.

La distribución de Poisson se utiliza para modelar el número de siniestros que ocurren en un intervalo de tiempo fijo. Aunque la distribución de Poisson es típicamente discreta, los modelos GLM le permiten tomar valores fraccionarios. Esta característica es útil al modelar la frecuencia cuando el número de siniestros de debe dividir entre el tiempo expuesto o la prima.

Cuando la distribución de Poisson tiene sobre-dispersión; es decir, que adicional a la varianza propia del proceso Poisson, existe también una varianza en la media del proceso, se puede utilizar una distribución Binomial Negativa.

Para modelar la severidad de los siniestros se utilizan principalmente las distribuciones gamma y Gausiana inversa. La distribución gamma es sesgada a la derecha, con un pico agudo, una larga cola a la derecha, y con límite inferior en 0. Como estas características tienden a presentarse en las distribuciones empíricas de la severidad de los siniestros, la distribución gamma se ajusta de forma natural (y de hecho es la distribución más utilizada) para modelizar la severidad en los GLM's (Goldburd, Khare, Tevet, & Guller, 2020)

La distribución gaussiana inversa, al igual que la distribución gamma, es sesgada a la derecha con límite inferior en 0 pero con un pico más pronunciado y una cola más larga. Por lo anterior, esta distribución es más adecuada en situaciones en las que se espera que haya una asimetría más pronunciada en la distribución.

2. Construir directamente modelos de prima pura utilizando la distribución de Tweedie

La distribución Tweedie tiene la característica de que su función de distribución de probabilidad tiene la mayor parte de su masa en 0 y la masa restante sesgada a la derecha; así mismo, además de los parámetros $\mu$ y $\emptyset$ de la familia exponencial estándar, la distribución Tweedie introduce un tercer parámetro, *p*, denominado parámetro de potencia, el cual puede tomar cualquier número real excepto los del intervalo 0 a 1 (0 y 1 son valores válidos).

Una característica de la distribución Tweedie es que varias de las otras distribuciones de la familia exponencial son casos especiales, dependiendo del valor de *p*. Por ejemplo:

- Una distribución Tweedie con $p = 0$ corresponde a una distribución normal.
- Una distribución Tweedie con $p = 1$ corresponde a una distribución Poisson.
- Una distribución Tweedie con $p = 2$ corresponde a una distribución gamma.
- Una distribución Tweedie con $p = 3$ corresponde a una distribución gaussiana inversa. (Dunn & Smyth, 2023)

Dado lo anterior, para modelar la prima pura de riesgo basta con escoger algún valor de *p* entre 1 (distribución Poisson – útil para modelar la frecuencia) y 2 (distribución gamma – útil para modelar la severidad). Con valores de *p* entre 1 y 2, Tweedie se convierte en una buena combinación de las distribuciones Poisson y gamma, que es ideal para modelar la prima pura o el índice de siniestralidad, es decir, los efectos combinados de la frecuencia y la severidad.

Finalmente, con el fin de obtener suficientes modelos que permitieran comparar el impacto de las variables obtenidas desde los dispositivos IoT, se construyeron los siguientes modelos de Frecuencia, Severidad y Prima pura:

- Modelo base: Construido únicamente con las variables categorizadas como tradicionales (insured.age.h, car.age.h, Insured.sex, Marital y years.noclaims.y.h)

- Modelo IoT 1: A partir del modelo base, se incluyen las variables IoT (pct.drive.wkday.h, Accel.09miles, Brake.09miles, Brake.11miles, Left.turn.intensity10 y Right.turn.intensity10)
- Modelo IoT 2: Modelo construido únicamente con las variables categorizadas como IoT
- Modelo IoT 3: Modelo construido a partir del algoritmo StepAIC incluyendo tanto las variables tradicionales como las variables IoT.

### 4.9.1. Modelos de frecuencia:

La construcción del modelo de frecuencia base presentó una significancia importante de las variables tradicionales que explican la variable objetivo, en este caso el número de reclamos esperados para una póliza. La mayoría de las variables incluidas en el modelo resultaron importantes con significancias inferiores al 0.001. Así mismo, todas las particiones realizadas a las variables son significativamente diferentes entre ellas. Finalmente, el modelo base construido presentó un AIC de 9520.5, el cual sirvió de punto de comparación con los demás modelos construidos.

Los modelos IoT1 y IoT3 (con dirección both, backward y forward) arrojaron los mismos resultados en términos de la significancia de las variables y de los intervalos construidos. Si bien, se observó que las categorías “Insured.age.h(26,60]”, “Accel.09miles.h[1,5]”, “Accel.11miles.h[1,3]” y “”Brake.11miles.h1” no difieren estadísticamente de sus respectivas categorías base por tener un p-value > 0.1, no se realizó una recategorización de las variables debido a las diferencias en las tasas de reclamos encontradas en la sección anterior y a las significancias generales de dichas variables.

A su vez, se observó que la inclusión de las variables telemáticas permitió obtener un mejor modelo que el modelo base, teniendo en cuenta que el AIC fue de 9230, un valor inferior que el encontrado con el modelo base.

Respecto al modelo IoT2 que incluye solamente variables telemáticas sin tener en cuenta variables tradicionales, se presentó un mayor número de categorías de las variables con p-value superiores al 0.1 indicando (al igual que en el caso anterior) que estas categorías no difieren estadísticamente de sus respectivas categorías base, si bien, al revisar la significancia general de las variables se obtuvo p-values inferiores al 0.01 en todos los casos.

Finalmente, al revisar el AIC del modelo IoT2, se obtuvo un valor de 9607 el cual fue superior al AIC tanto del modelo base como de los modelos IoT1 y IoT3, indicando que las variables telemáticas mejoran la precisión del modelo de frecuencia siempre y cuando se utilicen en conjunto con las variables tradicionales.

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -4.859783   0.214595 -22.646  < 2e-16 ***
insured.age.h(26,60]         -0.192803   0.121652  -1.585 0.112995
insured.age.h(60,100]        -0.804457   0.161955  -4.967 6.79e-07 ***
car.age.h(5,9]               -0.466636   0.069454  -6.719 1.83e-11 ***
car.age.h(9,19]              -1.607943   0.147275 -10.918  < 2e-16 ***
Insured.sexMale               0.176276   0.060874   2.896 0.003783 **
MaritalSingle                 0.377583   0.064482   5.856 4.75e-09 ***
years.noclaims.y.h(10,35]    -0.140271   0.082533  -1.700 0.089211 .
years.noclaims.y.h(35,55]    -0.248803   0.113128  -2.199 0.027856 *
years.noclaims.y.h(55,75]    -1.525165   0.465343  -3.278 0.001047 **
Pct.drive.wkday.h(0.7,0.8]    0.490441   0.083641   5.864 4.53e-09 ***
Pct.drive.wkday.h(0.8,1]      0.375780   0.103604   3.627 0.000287 ***
Accel.09miles.h[1, 5]         0.130072   0.079485   1.636 0.101750
Accel.09miles.h6+             0.725132   0.191392   3.789 0.000151 ***
Accel.11miles.h[1, 3]        -0.035536   0.090896  -0.391 0.695836
Accel.11miles.h4+            -0.624560   0.225359  -2.771 0.005582 **
Brake.09miles.h[1, 7]         0.627416   0.165196   3.798 0.000146 ***
Brake.09miles.h8+             1.046184   0.201863   5.183 2.19e-07 ***
Brake.11miles.h1             -0.005454   0.079538  -0.069 0.945334
Brake.11miles.h2+             0.258482   0.106632   2.424 0.015349 *
Left.turn.intensity10.h[1,10] 0.605497   0.117073   5.172 2.32e-07 ***
Left.turn.intensity10.h11+    0.679573   0.134369   5.058 4.25e-07 ***
Right.turn.intensity10.h[1,20] 0.372649  0.120480   3.093 0.001981 **
Right.turn.intensity10.h21+   0.352428   0.136293   2.586 0.009715 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 6961.5  on 35352  degrees of freedom
AIC: 9230.8

Number of Fisher Scoring iterations: 7
```

*Figura 26. Significancia estadística de los parámetros del modelo de frecuencia IoT 1. Elaboración propia*

### 4.9.2. Modelos de severidad:

La construcción del modelo de severidad base presentó una mayor dificultad debido a la baja significancia de las variables explicativas, en las que únicamente la variable “car.age” presenta una significancia inferior al 0.05. Dado que la variable objetivo en este modelo corresponde a una variable de tipo continuo, pues se está modelando el valor de los reclamos presentados, no se utilizan las categorías construidas de las variables continuas

sino las variables sin categorización. Este modelo inicial presentó un AIC de 5624, el cual sirvió de punto de comparación con los demás modelos construidos.

Los modelos IoT1 y IoT2 presentaron mayores dificultades en su construcción debido a que, en estos casos, tanto las variables tradicionales como telemáticas no presentan mayor significancia estadística generando AIC de 5621 y 36329, respectivamente. Es decir, modelos de comportamiento similar o “peor” que el modelo base.

El modelo IoT3 en dirección Both arrojó los mismos resultados en términos de las variables incluidas y de la significancia de estas que el modelo base, es decir, el algoritmo StepAIC incluyó la variable “Car.age” con una significancia similar a la encontrada en el modelo base aunque con un AIC de 36329 indicando que este modelos se no se ajusta tan bien a los datos como el modelo base.

Para el modelo IoT3 backward el algoritmo StepAIC construyó un modelo en el que las variables “Car.age”, “Insured.age”, “Brake.09miles” y “Brake.11miles” presentaron significancias menores al 0.05 y un AIC de 5613, indicando un ligero mejor ajuste que el modelo base construido.

Finalmente, para el modelo IoT3 forward el algoritmo StepAIC construyó un modelo en el que las variables “Car.age” e “Insured.age” presentaron significancias menores al 0.01 pero un AIC notablemente superior al del modelo base con un resultado de 31246, indicando que este modelo no se ajusta tan bien a los datos como el modelo base.

Teniendo en cuenta que el modelo de mejor AIC fue el modelo IoT3 backward el cual incluye las variables tradicionales “Car.age” y “Insured.age” y las variables telemáticas “Brake.09miles” y “Brake.11miles”, se podría pensar que las variables telemáticas aportan en los modelos de severidad construidos siempre y cuando se utilicen en conjunto con las variables tradicionales. Sin embargo, los niveles de significancia de estas variables no permiten asegurar dicha afirmación de manera categórica, por lo que se debe tener cuidado al momento de utilizar este modelo.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.59826    0.12431  12.857   <2e-16 ***
Car.age       -0.04086    0.01936  -2.110   0.0351 *
Insured.sexMale  0.12973    0.11910   1.089   0.2763
Brake.09miles    0.05735    0.02336   2.454   0.0143 *
Brake.11miles   -0.08837    0.04434  -1.993   0.0465 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.857526)

    Null deviance: 2240.3  on 1110  degrees of freedom
Residual deviance: 2192.1  on 1106  degrees of freedom
AIC: 5613

Number of Fisher Scoring iterations: 9
```

*Figura 27. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Backward". Elaboración propia*

### 4.9.3. Modelos Tweedie:

Para la construcción de los modelos Tweedie se generó un código iterativo en R que permitió generar y almacenar los resultados de los modelos construidos, variando el término *p* desde 1 hasta 2 y seleccionando aquel que generó el mejor ajuste. Los resultados presentados a continuación corresponden a aquellos obtenidos con el mejor parámetro *p*.

El modelo Tweedie base, al igual que en el caso del modelo de severidad, presentó dificultad en su construcción debido a la baja significancia de las variables explicativas, donde solo las variables "car.age" e "Insured.age" presentaron significancias inferiores al 0.05. Dado que la variable objetivo en este modelo corresponde a una variable de tipo continuo, pues se está modelando la prima pura de riesgo, no se utilizaron las categorías construidas de las variables continuas sino las variables sin categorización. Este modelo base presentó un AIC de 4366, el cual sirvió de punto de comparación con los demás modelos construidos.

En el modelo IoT1, se obtuvo que las variables tradicionales "car.age" e "Insured.age" y las variables telemáticas "Accel.09miles", "Accel.11.miles, "Brake.09miles" y "Brake.11miles" presentaron significancias inferiores al 0.05; sin embargo, el AIC obtenido fue de 4369 es decir, ligeramente superior al del modelo base.

En el modelo IoT2, se obtuvo que las variables "Accel.09miles", "Accel.11.miles, "Brake.09miles" y "Brake.11miles" presentaron significancias inferiores al 0.05; sin embargo, el AIC obtenido fue de 4376 es decir, superior al del modelo base.

Finalmente, para el modelo IoT3 el algoritmo StepAIC construyó un modelo en el que las variables tradicionales “Car.age” e “Insured.age” y las variables telemáticas “Accel.09miles”, “Accel.11.miles, “Brake.09miles” y “Brake.11miles” presentaron significancias menores al 0.05 con un AIC de 4362.

Teniendo en cuenta que el modelo de mejor AIC fue el modelo IoT3 el cual incluye las variables tradicionales “Car.age” y “Insured.age” y las variables telemáticas “Accel.09miles”, “Accel.11.miles, “Brake.09miles” y “Brake.11miles”, se podría pensar que las variables telemáticas aportan en los modelos de prima pura construidos siempre y cuando se utilicen en conjunto con las variables tradicionales.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.6531868  0.0129228 205.310  < 2e-16 ***
Insured.age    0.0005206  0.0002606   1.998 0.045948 *
Car.age       -0.0038815  0.0011311  -3.431 0.000622 ***
Accel.09miles -0.0036712  0.0020654  -1.777 0.075764 .
Accel.11miles  0.0128369  0.0045534   2.819 0.004900 **
Brake.09miles  0.0062209  0.0020395   3.050 0.002341 **
Brake.11miles -0.0164280  0.0050083  -3.280 0.001070 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 0.01433261)

    Null deviance: 17.898  on 1110  degrees of freedom
Residual deviance: 17.536  on 1104  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

*Figura 28. Significancia estadística de los parámetros del modelo Tweedie usando StepAIC en dirección Backward. Elaboración propia*

## 4.10. Resultados y comparación de los modelos

Una de las formas de comparar los modelos construidos es a través de los estadísticos AIC y del BIC.

El AIC (Akaike Information Criterion) se define como:

$$AIC = -2 \times log\ likelihood + 2p$$

Donde *p* es el número de parámetros en el modelo. En este caso un menor valor del AIC sugiere un mejor modelo.

El BIC (Bayesian Information Criterion) se define como:

$$BIC = -2 \times log\ likelihood + p\log(n)$$

Donde *p* es el número de parámetros del modelo y *n* es el número de datos en los que el modelo ajusta. De igual forma, un menor valor de este estadístico sugiere un mejor modelo

Dado que para la construcción de modelos se requiere una gran cantidad de datos, la penalización del estadístico BIC tiende a ser mayor que la penalización impuesta por el estadístico AIC por los parámetros utilizados.

En general, la mayoría de los paquetes estadísticos generan ambas medidas; sin embargo, en la práctica, los autores han descubierto que el AIC tiende a producir resultados más razonables, mientras que el BIC puede dar lugar a la exclusión de variables predictivas del modelo. (Goldburd, Khare, Tevet, & Guller, 2020)

Teniendo en cuenta lo anterior, se presenta a continuación la comparación de los modelos construidos:

### 4.10.1. Modelos de frecuencia

*Tabla 7. AIC de los modelos de frecuencia. Elaboración propia*

| Modelo | Variables | AIC |
|---|---|---|
| **IoT 1** | Insured.age, Car.age, Insured.sex, Marital, Years.noclaims.y, Pct.drive.wkday, Accel.09miles, Accel.11miles, Brake.09miles, Brake.11miles, Left.turn.intensity10, Right.turn.intensity10 | 9233 |
| **IoT 3 (both, forward, backward)** | Igual conjunto de variables que IoT 1 | 9233 |
| **Base** | Insured.age, Car.age, Insured.sex, Marital, Years.noclaims.y | 9522 |
| **IoT 2** | Accel.09miles, Accel.11miles, Brake.09miles, Brake.11miles, Left.turn.intensity10, Right.turn.intensity10 | 9609 |

### 4.10.2. Modelos de severidad

*Tabla 8. AIC de los modelos de severidad. Elaboración propia*

| Modelo | Variables | AIC |
|---|---|---|
| **IoT 3 backward** | Car.age, Insured.sex, Brake.09miles, Brake.11miles | 5613 |

| | | |
|---|---|---|
| **IoT 1** | Insured.age, Car.age, Insured.sex, Marital, Years.noclaims.y, Pct.drive.wkday, Accel.09miles, Accel.11miles, Brake.09miles, Brake.11miles, Left.turn.intensity10, Right.turn.intensity10 | 5621 |
| **Base** | Insured.age, Car.age, Insured.sex, Marital, Years.noclaims.y | 5625 |
| **IoT 3 forward** | Car.age, Insured.age | 31246 |
| **IoT 2** | Accel.09miles, Accel.11miles, Brake.09miles, Brake.11miles, Left.turn.intensity10, Right.turn.intensity10 | 36330 |
| **IoT 3 both** | Car.age | 36330 |

### 4.10.3. Modelos Tweedie

*Tabla 9. AIC de los modelos Tweedie. Elaboración propia*

| Modelo | Variables | AIC |
|---|---|---|
| **IoT 3 both** | Insured.age, Car.age, Accel.09miles, Accel.11miles, Brake.09miles, Brake.11miles | 4362 |
| **Base** | Insured.age, Car.age, Insured.sex, Marital, Years.noclaims.y | 4366 |
| **IoT 1** | Variables tradicionales + variables telemáticas | 4369 |
| **IoT 2** | Accel.09miles, Accel.11miles, Brake.09miles, Brake.11miles, Left.turn.intensity10, Right.turn.intensity10 | 4376 |

Los resultados de la Tabla 7, Tabla 8 y Tabla 9 muestran que a pesar de que la inclusión de variables telemáticas en los modelos generó una disminución en el valor del AIC, lo cual sugiere una mejora en la calidad de los modelos, esta mejora no fue tan significativa como se esperaba al principio del proyecto. En particular, los modelos de severidad y Tweedie no presentan mejoras tan significativas como en el modelo de frecuencia.

Una posible explicación para esta situación podría ser la complejidad de la relación entre las variables telemáticas y el comportamiento de los conductores en Colombia. Aunque se utilizaron variables telemáticas simuladas basadas en la experiencia de seguros de autos de Canadá y, por ende, en estilos de conducción similares, es posible que existan diferencias significativas entre los hábitos de conducción de los conductores canadienses y colombianos, afectando la capacidad de las variables telemáticas para mejorar los modelos.

En consecuencia, aunque los modelos con variables telemáticas resultaron en una mejora en el valor del AIC en general, se debe tener precaución al interpretar estos resultados y se recomienda explorar más a fondo la relación entre las variables telemáticas y el comportamiento de los conductores en Colombia para mejorar la construcción de modelos de severidad (principalmente) más precisos y efectivos.

# 5. Conclusiones y recomendaciones

La industria aseguradora está descubriendo cómo los avances en la tecnología de los dispositivos IoT y el procesamiento de grandes cantidades de datos pueden ayudar a calcular el perfil de riesgo de sus clientes casi en tiempo real. Esto les permite crear productos cada vez más personalizados y adaptados a las necesidades de los asegurados. En particular, en el caso de los seguros de automóviles, las variables telemáticas extraídas de los dispositivos IoT pueden recopilar información en tiempo real sobre el comportamiento y habilidades del conductor, lo que permite a las aseguradoras analizar sus patrones de conducción y calcular con mayor precisión las probabilidades de tener un siniestro.

En este contexto, este proyecto tuvo como principal objetivo la construcción de modelos de prima pura de riesgo mediante la modelación de la Frecuencia y la Severidad y de modelos Tweedie incluyendo variables telemáticas para determinar su aporte. Para ello, se recopiló información sobre pólizas de autos expedidas en Colombia entre 2020 y 2022, incluyendo variables telemáticas simuladas a partir del comportamiento de asegurados en Canadá, provenientes de la publicación "Synthetic Dataset Generation of Driver Telematics" (So, Boucher, & Valdez, 2021).

Una vez construidas las bases de datos, se construyeron múltiples modelos de tarifación utilizando diferentes conjuntos de variables tanto tradicionales como telemáticas mediante modelos GLM y utilizando el algoritmo StepAIC para la selección de las variables más significantes y así validar el ajuste a los datos.

En particular, la revisión del estadístico AIC como método de comparación entre los modelos construidos mostró que la inclusión de variables telemáticas provenientes de dispositivos IoT ayudó a disminuir el valor resultante del AIC en comparación con los resultados obtenidos en los modelos base de frecuencia y, en menor medida, en los modelos de severidad y Tweedie, indicando que estas variables no tradicionales pueden aportar en la construcción de modelos más precisos, eficientes y justos y, por lo tanto, en una mayor rentabilidad para las compañías de seguros. Es importante tener en cuenta que estos resultados, principalmente el de los modelos de severidad y Tweedie, pueden estar afectados por las diferencias entre las poblaciones con las que se construyó la base para la modelación, como se explicó anteriormente.

Así mismo, la inclusión de variables telemáticas en la tarificación de seguros de autos puede tener un impacto desigual en los asegurados debido a que los conductores que tengan hábitos de conducción más seguros pueden verse beneficiados por una tarificación más personalizada y, por lo tanto, por una prima más baja. Por el contrario, los conductores con hábitos de conducción más arriesgados pueden ver su prima aumentada lo cual puede ser percibido como injusto por algunos asegurados.

De igual forma, a lo largo del desarrollo del proyecto se encontró con la dificultad de no contar con un mayor número de variables tradicionales que hubiesen podido ser tenidas en cuenta para la construcción de los modelos base como, por ejemplo: marca y línea del auto, cuidad de circulación, tipo de uso, entre otras, que podrían ayudar a desarrollar mejores modelos base de comparación. Otra dificultad presentada corresponde a que la base unificada con la que se desarrollaron los modelos no contiene una cantidad significativa de reclamos que dificulta, a su vez, la construcción de modelos más precisos o el uso de métodos alternativos de ajuste de los parámetros como k-folds.

Para obtener resultados más precisos y conclusiones más robustas se recomienda, en futuros proyectos, desarrollar los siguientes puntos:

- Obtener bases de datos reales con variables telemáticas de pólizas de autos expedidas en Colombia o de países con comportamientos y/o hábitos de conducción similares a los colombianos.
- Realizar ingeniería de características con las variables tradicionales y telemáticas para descubrir nuevas relaciones que potencien la construcción de los modelos. La ingeniería de características (feature engineering en inglés), es el proceso de selección, extracción y transformación de variables o atributos (características) relevantes a partir de los datos en bruto. Una buena ingeniería de características puede mejorar significativamente la precisión y eficiencia de los modelos construidos, ya que permite una mejor representación de los datos y una mayor capacidad para capturar patrones complejos y relaciones entre variables.
- Finalmente, dada la cantidad de variables telemáticas que se pueden obtener desde los dispositivos IoT se recomienda construir modelos ya sea de frecuencia / severidad o de prima pura de riesgo utilizando técnicas diferentes a los modelos GLM como, por ejemplo: algoritmos de machine learning como Random Forests, XGBoost o incluso redes neuronales, que puedan aprovechar la cantidad de información obtenida desde los dispositivos IoT y no se vean demasiado limitados por los diferentes tipos de datos.

## 6. Referencias bibliográficas


Amazon. (Mayo de 2021). *Amazon*. Obtenido de https://www.amazon.com/-/es/dp/B07XJ8C8F5/ref=s9_acsd_al_bw_c2_x_2_t?pf_rd_m=ATVPDKIKX0DER&pf_rd_s=merchandised-search-1&pf_rd_r=6XMVK8DYYSS6G3C722PG&pf_rd_t=101&pf_rd_p=eb2cc9ea-8f02-4bab-9f1e-c5d23523b604&pf_rd_i=9818047011

Anderson, C. (2023). *University of Illinois.* Obtenido de Department of Educational Psychology: https://education.illinois.edu/docs/default-source/carolyn-anderson/edpsy589/lectures/4_glm/4glm_1_beamer_post.pdf?sfvrsn=e35ac12_12

Deloitte. (s.f.). *IoT - Internet Of Things*. Obtenido de https://www2.deloitte.com/es/es/pages/technology/articles/IoT-internet-of-things.html

Deloitte. (s.f.). *La transformación de las compañías de seguros en la era digital*. Obtenido de https://www2.deloitte.com/uy/es/pages/strategy-operations/articles/La-transformacion-de-las-companias-de-seguros-en-la-era-digital.html

Dunn, P. K., & Smyth, G. K. (2023). *Generalized Linear Models With Examples in R.* Springer Texts in Statistics.

Falabella. (s.f.). *Seguro x km*. Obtenido de https://web.segurosfalabella.com/cl/seguros-de-auto/x-kilometro/

FASECOLDA. (s.f.). *Acerca de la Guía de Valores*. Obtenido de https://fasecolda.com/ramos/automoviles/guia-de-valores/

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2020). *GENERALIZED LINEAR MODELS FOR INSURANCE RATING.*

Google. (Mayo de 2021). Obtenido de https://store.google.com/us/product/nest_audio?hl=en-US

Internet of Business. (s.f.). *10 real-life examples of IoT in insurance*. Obtenido de https://internetofbusiness.com/10-examples-iot-insurance/

Ministerio de salud y protección social. (02 de 05 de 2023). *www.minsalud.gov.co*. Obtenido de https://www.minsalud.gov.co/proteccionsocial/Paginas/cicloVida.aspx#:~:text=La%20siguiente%20clasificaci%C3%B3n%20es%20un,(60%20a%C3%B1os%20y%20m%C3%A1s).

Oracle. (s.f.). *¿Que es Big Data?* Obtenido de https://www.oracle.com/co/big-data/what-is-big-data/

Rose, K., Eldrige, S., & Chapin, L. (Octubre de 2015). *La internet de las cosas - Una breve reseña.* Obtenido de Internet Society: https://www.internetsociety.org/wp-content/uploads/2017/09/report-InternetOfThings-20160817-es-1.pdf

Seguros Bolivar. (s.f.). *Seguro de carro por recorridos*. Obtenido de https://www.segurosbolivar.com/seguro-de-autos-por-recorridos

Semana. (Abril de 2023). *Los requisitos que debe tener en cuenta para comprar un carro a crédito en Colombia*. Obtenido de https://www.semana.com/finanzas/consumo-inteligente/articulo/los-requisitos-que-debe-tener-en-cuenta-para-comprar-un-carro-a-credito-en-colombia/202314/

Shlens, J. (25 de Marzo de 2003). *Princeton.* Obtenido de A tutorial on principal component analysis. Derivation, Discussion and Singular Value Decomposition: https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf

So, B., Boucher, J.-P., & Valdez, E. (2021). arxiv.org. *Synthetic Dataset Generation of Driver Telematics*. Obtenido de https://arxiv.org/abs/2102.00252

SUMA movil. (Febrero de 2023). *El negocio M2M/IoT en Colombia cuenta con un nuevo aliado*. Obtenido de https://sumamovil.com.co/el-negocio-m2m-iot-en-colombia-cuenta-con-un-nuevo-aliado/

Williams, L. J., & Abdi, H. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 433-459.

# Anexos

## Anexo 1 Factores de desarrollo de pólizas reales

| mes_aviso | 0_1 | 1_2 | 2_3 | 3_4 | 4_5 | 5_6 | 6_7 | 7_8 | 8_9 | 9_10 | 10_11 | 11_12 | 12_13 | 13_14 | 14_15 | 15_16 | 16_17 | 17_18 | 18_19 | 19_20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2020_01 | 3,4597 | 1,2757 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0186 | 1,0000 | 1,0464 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0105 |
| 2020_02 | 1,7481 | 1,0483 | 1,0465 | 1,1493 | 1,0375 | 1,0429 | 1,0596 | 1,0355 | 1,0000 | 1,0055 | 1,0000 | 1,0000 | 1,0237 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |
| 2020_03 | 1,0773 | 1,1567 | 1,0367 | 1,0372 | 1,0183 | 1,0028 | 1,0000 | 1,0057 | 1,0087 | 1,0000 | 1,0000 | 1,0169 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |
| 2020_04 | 1,4088 | 1,1799 | 1,0000 | 1,0000 | 1,0178 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |
| 2020_05 | 1,1278 | 1,0719 | 1,0854 | 1,0060 | 1,0023 | 1,0000 | 1,0000 | 1,0128 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0631 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |
| 2020_06 | 1,0043 | 1,0107 | 1,0034 | 1,0010 | 1,0000 | 1,0038 | 1,0061 | 1,0017 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |
| 2020_07 | 1,0937 | 1,0175 | 1,0019 | 1,0058 | 1,0043 | 1,0082 | 1,0000 | 1,0034 | 1,0000 | 1,0000 | 1,0013 | 1,0000 | 1,0024 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0073 | 1,0000 | 1,0000 |
| 2020_08 | 1,0151 | 1,0067 | 1,0041 | 1,0084 | 1,0045 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | |
| 2020_09 | 1,0284 | 1,0096 | 1,0269 | 1,0053 | 1,0016 | 1,0000 | 1,0089 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0007 | 1,0000 | 1,0000 | 1,0000 | | |
| 2020_10 | 1,0652 | 1,0093 | 1,0014 | 1,0011 | 1,0073 | 1,0007 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0007 | 1,0000 | 1,0000 | | | |
| 2020_11 | 1,0224 | 1,0292 | 1,0020 | 1,0128 | 1,0039 | 1,0000 | 1,0000 | 1,0038 | 1,0002 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | | | | |
| 2020_12 | 1,0075 | 1,0032 | 1,0198 | 1,0031 | 1,0007 | 1,0000 | 1,0000 | 1,0089 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0003 | 1,0000 | 1,0000 | | | | | |
| 2021_01 | 1,0027 | 1,0130 | 1,0115 | 1,0014 | 1,0000 | 1,0061 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | | | | | | |
| 2021_02 | 1,0481 | 1,0061 | 1,0083 | 1,0001 | 1,0007 | 1,0000 | 1,0000 | 1,0000 | 1,0015 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | | | | | | | |
| 2021_03 | 1,0106 | 1,0017 | 1,0014 | 1,0103 | 1,0000 | 1,0009 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | | | | | | | | |
| 2021_04 | 1,0007 | 1,0013 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | | | | | | | | | |
| 2021_05 | 1,0000 | 1,0139 | 1,0000 | 1,0006 | 1,0000 | 1,0000 | 1,0000 | 1,0002 | 1,0027 | 1,0000 | | | | | | | | | | |
| 2021_06 | 1,0258 | 1,0182 | 1,0038 | 1,0043 | 1,0000 | 1,0010 | 1,0009 | 1,0000 | 1,0000 | | | | | | | | | | | |
| 2021_07 | 1,0143 | 1,0079 | 1,0188 | 1,0016 | 1,0011 | 1,0011 | 1,0129 | 1,0009 | | | | | | | | | | | | |
| 2021_08 | 1,0101 | 1,0125 | 1,0008 | 1,0004 | 1,0001 | 1,0005 | 1,0000 | | | | | | | | | | | | | |
| 2021_09 | 1,0041 | 1,0124 | 1,0046 | 1,0016 | 1,0025 | 1,0000 | | | | | | | | | | | | | | |
| 2021_10 | 1,0267 | 1,0175 | 1,0138 | 1,0048 | 1,0000 | | | | | | | | | | | | | | | |
| 2021_11 | 1,0021 | 1,0053 | 1,0105 | 1,0017 | | | | | | | | | | | | | | | | |
| 2021_12 | 1,0001 | 1,0002 | 1,0000 | | | | | | | | | | | | | | | | | |
| 2022_01 | 1,0056 | 1,0009 | | | | | | | | | | | | | | | | | | |
| 2022_02 | 1,0000 | | | | | | | | | | | | | | | | | | | |
| 2022_03 | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | |
| Promedio | 1,1619 | 1,0372 | 1,0126 | 1,0112 | 1,0047 | 1,0032 | 1,0053 | 1,0038 | 1,0033 | 1,0003 | 1,0001 | 1,0011 | 1,0019 | 1,0049 | 1,0001 | 1,0000 | 1,0000 | 1,0008 | 1,0000 | 1,0015 |

*Figura 29. Factores de desarrollo de pólizas reales*

## Anexo 2 Resultados de los modelos de frecuencia

- Modelo base:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.86448    0.11874 -24.125  < 2e-16 ***
insured.age.h(26,60]      -0.21399    0.12107  -1.767  0.07715 .
insured.age.h(60,100]     -0.87922    0.16006  -5.493 3.95e-08 ***
car.age.h(5,9]            -0.36816    0.06823  -5.396 6.83e-08 ***
car.age.h(9,19]           -1.57166    0.14606 -10.760  < 2e-16 ***
Insured.sexMale            0.19169    0.06058   3.164  0.00156 **
MaritalSingle              0.40264    0.06417   6.275 3.51e-10 ***
years.noclaims.y.h(10,35] -0.21768    0.08155  -2.669  0.00760 **
years.noclaims.y.h(35,55] -0.46920    0.11036  -4.251 2.12e-05 ***
years.noclaims.y.h(55,75] -1.84686    0.46446  -3.976 7.00e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 7279.1  on 35366  degrees of freedom
AIC: 9520.5

Number of Fisher Scoring iterations: 7
```

*Figura 30. Significancia estadística de los parámetros del modelo de frecuencia base. Elaboración propia*

```
> Anova(modelo.frec_base, type=3)
Analysis of Deviance Table (Type III tests)

Response: NB_Claim
                  LR Chisq Df Pr(>Chisq)
insured.age.h       45.352  2  1.419e-10 ***
car.age.h          190.728  2  < 2.2e-16 ***
Insured.sex         10.009  1   0.001558 **
Marital             37.909  1  7.411e-10 ***
years.noclaims.y.h  34.510  3  1.546e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figura 31 Tabla ANOVA del modelo de frecuencia base*

- Modelo IoT 1:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.859783   0.214595 -22.646  < 2e-16 ***
insured.age.h(26,60]     -0.192803   0.121652  -1.585 0.112995
insured.age.h(60,100]    -0.804457   0.161955  -4.967 6.79e-07 ***
car.age.h(5,9]           -0.466636   0.069454  -6.719 1.83e-11 ***
car.age.h(9,19]          -1.607943   0.147275 -10.918  < 2e-16 ***
Insured.sexMale           0.176276   0.060874   2.896 0.003783 **
MaritalSingle             0.377583   0.064482   5.856 4.75e-09 ***
years.noclaims.y.h(10,35] -0.140271   0.082533  -1.700 0.089211 .
years.noclaims.y.h(35,55] -0.248803   0.113128  -2.199 0.027856 *
years.noclaims.y.h(55,75] -1.525165   0.465343  -3.278 0.001047 **
Pct.drive.wkday.h(0.7,0.8] 0.490441   0.083641   5.864 4.53e-09 ***
Pct.drive.wkday.h(0.8,1]  0.375780   0.103604   3.627 0.000287 ***
Accel.09miles.h[1, 5]     0.130072   0.079485   1.636 0.101750
Accel.09miles.h6+         0.725132   0.191392   3.789 0.000151 ***
Accel.11miles.h[1, 3]    -0.035536   0.090896  -0.391 0.695836
Accel.11miles.h4+        -0.624560   0.225359  -2.771 0.005582 **
Brake.09miles.h[1, 7]     0.627416   0.165196   3.798 0.000146 ***
Brake.09miles.h8+         1.046184   0.201863   5.183 2.19e-07 ***
Brake.11miles.h1         -0.005454   0.079538  -0.069 0.945334
Brake.11miles.h2+         0.258482   0.106632   2.424 0.015349 *
Left.turn.intensity10.h[1,10] 0.605497   0.117073   5.172 2.32e-07 ***
Left.turn.intensity10.h11+ 0.679573   0.134369   5.058 4.25e-07 ***
Right.turn.intensity10.h[1,20] 0.372649   0.120480   3.093 0.001981 **
Right.turn.intensity10.h21+ 0.352428   0.136293   2.586 0.009715 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 6961.5  on 35352  degrees of freedom
AIC: 9230.8

Number of Fisher Scoring iterations: 7
```

*Figura 32. Significancia estadística de los parámetros del modelo de frecuencia IoT 1. Elaboración propia*

```
> Anova(modelo.frec_IoT_1, type=3)
Analysis of Deviance Table (Type III tests)

Response: NB_Claim
                     LR Chisq Df Pr(>Chisq)
insured.age.h          36.950  2  9.470e-09 ***
car.age.h             201.664  2  < 2.2e-16 ***
Insured.sex             8.353  1  0.0038511 **
Marital                32.979  1  9.314e-09 ***
years.noclaims.y.h     17.714  3  0.0005038 ***
Pct.drive.wkday.h      37.589  2  6.882e-09 ***
Accel.09miles.h        12.755  2  0.0016994 **
Accel.11miles.h         7.725  2  0.0210125 *
Brake.09miles.h        29.408  2  4.113e-07 ***
Brake.11miles.h         8.509  2  0.0141993 *
Left.turn.intensity10.h 30.900  2  1.950e-07 ***
Right.turn.intensity10.h 10.601  2  0.0049903 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figura 33 Tabla ANOVA del modelo de frecuencia IoT1*

- Modelo IoT 2:

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -5.46596    0.17871 -30.585  < 2e-16 ***
Pct.drive.wkday.h(0.7,0.8]  0.38899    0.08315   4.678 2.90e-06 ***
Pct.drive.wkday.h(0.8,1]    0.12946    0.10254   1.262 0.206771
Accel.09miles.h[1, 5]       0.11106    0.07944   1.398 0.162084
Accel.09miles.h6+           0.69952    0.18910   3.699 0.000216 ***
Accel.11miles.h[1, 3]      -0.09610    0.09049  -1.062 0.288238
Accel.11miles.h4+          -0.80369    0.22299  -3.604 0.000313 ***
Brake.09miles.h[1, 7]       0.69160    0.16501   4.191 2.77e-05 ***
Brake.09miles.h8+           1.13043    0.20117   5.619 1.92e-08 ***
Brake.11miles.h1            0.03517    0.07913   0.444 0.656687
Brake.11miles.h2+           0.30646    0.10539   2.908 0.003641 **
Left.turn.intensity10.h[1,10]  0.69394    0.11734   5.914 3.34e-09 ***
Left.turn.intensity10.h11+  0.82180    0.13234   6.210 5.31e-10 ***
Right.turn.intensity10.h[1,20]  0.44130    0.12092   3.650 0.000263 ***
Right.turn.intensity10.h21+  0.42229    0.13490   3.130 0.001746 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 7355.8  on 35361  degrees of freedom
AIC: 9607.2
```

*Figura 34. Significancia estadística de los parámetros del modelo de frecuencia IoT 2. Elaboración propia*

```
> Anova(modelo.frec_IoT_2, type=3)
Analysis of Deviance Table (Type III tests)

Response: NB_Claim
                      LR Chisq Df Pr(>Chisq)
Pct.drive.wkday.h       28.648  2  6.015e-07 ***
Accel.09miles.h         12.093  2  0.0023664 **
Accel.11miles.h         12.254  2  0.0021836 **
Brake.09miles.h         34.693  2  2.928e-08 ***
Brake.11miles.h         10.133  2  0.0063057 **
Left.turn.intensity10.h 44.310  2  2.388e-10 ***
Right.turn.intensity10.h 14.873  2  0.0005894 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 35 Tabla ANOVA del modelo de frecuencia IoT2*

- Modelo IoT 3:
  - Step AIC dirección “Both”

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.859783   0.214595 -22.646  < 2e-16 ***
Left.turn.intensity10.h[1,10]   0.605497   0.117073   5.172 2.32e-07 ***
Left.turn.intensity10.h11+      0.679573   0.134369   5.058 4.25e-07 ***
car.age.h(5,9]                 -0.466636   0.069454  -6.719 1.83e-11 ***
car.age.h(9,19]                -1.607943   0.147275 -10.918  < 2e-16 ***
insured.age.h(26,60]           -0.192803   0.121652  -1.585 0.112995
insured.age.h(60,100]          -0.804457   0.161955  -4.967 6.79e-07 ***
Brake.09miles.h[1, 7]           0.627416   0.165196   3.798 0.000146 ***
Brake.09miles.h8+               1.046184   0.201863   5.183 2.19e-07 ***
MaritalSingle                   0.377583   0.064482   5.856 4.75e-09 ***
Pct.drive.wkday.h(0.7,0.8]      0.490441   0.083641   5.864 4.53e-09 ***
Pct.drive.wkday.h(0.8,1]        0.375780   0.103604   3.627 0.000287 ***
years.noclaims.y.h(10,35]      -0.140271   0.082533  -1.700 0.089211 .
years.noclaims.y.h(35,55]      -0.248803   0.113128  -2.199 0.027856 *
years.noclaims.y.h(55,75]      -1.525165   0.465343  -3.278 0.001047 **
Accel.09miles.h[1, 5]           0.130072   0.079485   1.636 0.101750
Accel.09miles.h6+               0.725132   0.191392   3.789 0.000151 ***
Insured.sexMale                 0.176276   0.060874   2.896 0.003783 **
Right.turn.intensity10.h[1,20]  0.372649   0.120480   3.093 0.001981 **
Right.turn.intensity10.h21+     0.352428   0.136293   2.586 0.009715 **
Brake.11miles.h1               -0.005454   0.079538  -0.069 0.945334
Brake.11miles.h2+               0.258482   0.106632   2.424 0.015349 *
Accel.11miles.h[1, 3]          -0.035536   0.090896  -0.391 0.695836
Accel.11miles.h4+              -0.624560   0.225359  -2.771 0.005582 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 6961.5  on 35352  degrees of freedom
AIC: 9230.8

Number of Fisher Scoring iterations: 7
```

*Figura 36. Significancia estadística de los parámetros del modelo de frecuencia IoT 3 usando StepAIC en dirección "Both". Elaboración propia*

```
> Anova(modelo.frec_IoT_3_both, type=3)
Analysis of Deviance Table (Type III tests)

Response: NB_Claim
                       LR Chisq Df Pr(>Chisq)
Left.turn.intensity10.h   30.900  2  1.950e-07 ***
car.age.h                201.664  2  < 2.2e-16 ***
insured.age.h             36.950  2  9.470e-09 ***
Brake.09miles.h           29.408  2  4.113e-07 ***
Marital                   32.979  1  9.314e-09 ***
Pct.drive.wkday.h         37.589  2  6.882e-09 ***
years.noclaims.y.h        17.714  3  0.0005038 ***
Accel.09miles.h           12.755  2  0.0016994 **
Right.turn.intensity10.h  10.601  2  0.0049903 **
Insured.sex                8.353  1  0.0038511 **
Brake.11miles.h            8.509  2  0.0141993 *
Accel.11miles.h            7.725  2  0.0210125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 37 Tabla ANOVA del modelo de frecuencia IoT3 en dirección Both*

- Step AIC dirección “Backward”

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.859783   0.214595 -22.646  < 2e-16 ***
insured.age.h(26,60]     -0.192803   0.121652  -1.585 0.112995
insured.age.h(60,100]    -0.804457   0.161955  -4.967 6.79e-07 ***
car.age.h(5,9]           -0.466636   0.069454  -6.719 1.83e-11 ***
car.age.h(9,19]          -1.607943   0.147275 -10.918  < 2e-16 ***
Insured.sexMale           0.176276   0.060874   2.896 0.003783 **
MaritalSingle             0.377583   0.064482   5.856 4.75e-09 ***
years.noclaims.y.h(10,35] -0.140271   0.082533  -1.700 0.089211 .
years.noclaims.y.h(35,55] -0.248803   0.113128  -2.199 0.027856 *
years.noclaims.y.h(55,75] -1.525165   0.465343  -3.278 0.001047 **
Pct.drive.wkday.h(0.7,0.8] 0.490441   0.083641   5.864 4.53e-09 ***
Pct.drive.wkday.h(0.8,1]  0.375780   0.103604   3.627 0.000287 ***
Accel.09miles.h[1, 5]     0.130072   0.079485   1.636 0.101750
Accel.09miles.h6+         0.725132   0.191392   3.789 0.000151 ***
Accel.11miles.h[1, 3]    -0.035536   0.090896  -0.391 0.695836
Accel.11miles.h4+        -0.624560   0.225359  -2.771 0.005582 **
Brake.09miles.h[1, 7]     0.627416   0.165196   3.798 0.000146 ***
Brake.09miles.h8+         1.046184   0.201863   5.183 2.19e-07 ***
Brake.11miles.h1         -0.005454   0.079538  -0.069 0.945334
Brake.11miles.h2+         0.258482   0.106632   2.424 0.015349 *
Left.turn.intensity10.h[1,10] 0.605497   0.117073   5.172 2.32e-07 ***
Left.turn.intensity10.h11+ 0.679573   0.134369   5.058 4.25e-07 ***
Right.turn.intensity10.h[1,20] 0.372649   0.120480   3.093 0.001981 **
Right.turn.intensity10.h21+ 0.352428   0.136293   2.586 0.009715 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 6961.5  on 35352  degrees of freedom
AIC: 9230.8
```

*Figura 38. Significancia estadística de los parámetros del modelo de frecuencia IoT 3 usando StepAIC en dirección "Backward". Elaboración propia*

```
> Anova(modelo.frec_IoT_3_bw, type=3)
Analysis of Deviance Table (Type III tests)

Response: NB_Claim
                      LR Chisq Df Pr(>Chisq)
insured.age.h           36.950  2  9.470e-09 ***
car.age.h              201.664  2  < 2.2e-16 ***
Insured.sex              8.353  1  0.0038511 **
Marital                 32.979  1  9.314e-09 ***
years.noclaims.y.h      17.714  3  0.0005038 ***
Pct.drive.wkday.h       37.589  2  6.882e-09 ***
Accel.09miles.h         12.755  2  0.0016994 **
Accel.11miles.h          7.725  2  0.0210125 *
Brake.09miles.h         29.408  2  4.113e-07 ***
Brake.11miles.h          8.509  2  0.0141993 *
Left.turn.intensity10.h 30.900  2  1.950e-07 ***
Right.turn.intensity10.h 10.601  2  0.0049903 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 39 Tabla ANOVA del modelo de frecuencia IoT3 en dirección Backward*

- Step AIC dirección “Forward”

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.859783   0.214595 -22.646  < 2e-16 ***
Left.turn.intensity10.h[1,10]   0.605497   0.117073   5.172 2.32e-07 ***
Left.turn.intensity10.h11+      0.679573   0.134369   5.058 4.25e-07 ***
car.age.h(5,9]                 -0.466636   0.069454  -6.719 1.83e-11 ***
car.age.h(9,19]                -1.607943   0.147275 -10.918  < 2e-16 ***
insured.age.h(26,60]           -0.192803   0.121652  -1.585 0.112995
insured.age.h(60,100]          -0.804457   0.161955  -4.967 6.79e-07 ***
Brake.09miles.h[1, 7]           0.627416   0.165196   3.798 0.000146 ***
Brake.09miles.h8+               1.046184   0.201863   5.183 2.19e-07 ***
MaritalSingle                   0.377583   0.064482   5.856 4.75e-09 ***
Pct.drive.wkday.h(0.7,0.8]      0.490441   0.083641   5.864 4.53e-09 ***
Pct.drive.wkday.h(0.8,1]        0.375780   0.103604   3.627 0.000287 ***
years.noclaims.y.h(10,35]      -0.140271   0.082533  -1.700 0.089211 .
years.noclaims.y.h(35,55]      -0.248803   0.113128  -2.199 0.027856 *
years.noclaims.y.h(55,75]      -1.525165   0.465343  -3.278 0.001047 **
Accel.09miles.h[1, 5]           0.130072   0.079485   1.636 0.101750
Accel.09miles.h6+               0.725132   0.191392   3.789 0.000151 ***
Insured.sexMale                 0.176276   0.060874   2.896 0.003783 **
Right.turn.intensity10.h[1,20]  0.372649   0.120480   3.093 0.001981 **
Right.turn.intensity10.h21+     0.352428   0.136293   2.586 0.009715 **
Brake.11miles.h1               -0.005454   0.079538  -0.069 0.945334
Brake.11miles.h2+               0.258482   0.106632   2.424 0.015349 *
Accel.11miles.h[1, 3]          -0.035536   0.090896  -0.391 0.695836
Accel.11miles.h4+              -0.624560   0.225359  -2.771 0.005582 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7762.4  on 35375  degrees of freedom
Residual deviance: 6961.5  on 35352  degrees of freedom
AIC: 9230.8
```

*Figura 40. Significancia estadística de los parámetros del modelo de frecuencia IoT 3 usando StepAIC en dirección "Forward". Elaboración propia*

```
> Anova(modelo.frec_IoT_3_fw, type=3)
Analysis of Deviance Table (Type III tests)

Response: NB_Claim
                        LR Chisq Df Pr(>Chisq)
Left.turn.intensity10.h   30.900  2  1.950e-07 ***
car.age.h                201.664  2  < 2.2e-16 ***
insured.age.h             36.950  2  9.470e-09 ***
Brake.09miles.h           29.408  2  4.113e-07 ***
Marital                   32.979  1  9.314e-09 ***
Pct.drive.wkday.h         37.589  2  6.882e-09 ***
years.noclaims.y.h        17.714  3  0.0005038 ***
Accel.09miles.h           12.755  2  0.0016994 **
Right.turn.intensity10.h  10.601  2  0.0049903 **
Insured.sex                8.353  1  0.0038511 **
Brake.11miles.h            8.509  2  0.0141993 *
Accel.11miles.h            7.725  2  0.0210125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 41 Tabla ANOVA del modelo de frecuencia IoT3 en dirección Forward*

## Anexo 3 Resultados de los modelos de severidad

- Modelo base:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.375192   0.309652   4.441 9.85e-06 ***
Insured.age     0.009023   0.008210   1.099   0.2720
Car.age        -0.042334   0.019390  -2.183   0.0292 *
Insured.sexMale 0.126125   0.120478   1.047   0.2954
MaritalSingle   0.139310   0.139483   0.999   0.3181
Years.noclaims.y -0.006256   0.007471  -0.837   0.4025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.910143)

    Null deviance: 2240.3  on 1110  degrees of freedom
Residual deviance: 2207.6  on 1105  degrees of freedom
AIC: 5624.9
```

*Figura 42. Significancia estadística de los parámetros del modelo de severidad base. Elaboración propia*

```
> Anova(modelo.sev_base_gamma, type=3)
Analysis of Deviance Table (Type III tests)

Response: AMT_Claim.x/1e+06
               LR Chisq Df Pr(>Chisq)
Insured.age      1.0566  1     0.3040
Car.age          5.1759  1     0.0229 *
Insured.sex      1.0922  1     0.2960
Marital          0.9872  1     0.3204
Years.noclaims.y 0.6467  1     0.4213
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 43 Tabla ANOVA del modelo de severidad base*

- Modelo IoT 1:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.927e-01  7.874e-01   0.880   0.3792
Insured.age              6.281e-03  8.298e-03   0.757   0.4492
Car.age                 -4.301e-02  1.978e-02  -2.174   0.0299 *
Insured.sexMale          1.245e-01  1.212e-01   1.027   0.3046
MaritalSingle            1.430e-01  1.400e-01   1.021   0.3075
Years.noclaims.y        -2.321e-03  7.693e-03  -0.302   0.7629
Pct.drive.wkday          8.245e-01  9.907e-01   0.832   0.4055
Accel.09miles            7.508e-03  3.621e-02   0.207   0.8358
Accel.11miles            2.386e-02  7.965e-02   0.300   0.7646
Brake.09miles            6.715e-02  3.602e-02   1.864   0.0626 .
Brake.11miles           -1.304e-01  8.769e-02  -1.487   0.1374
Left.turn.intensity10    1.260e-07  3.170e-06   0.040   0.9683
Right.turn.intensity10   1.426e-06  3.831e-06   0.372   0.7098
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.900652)

    Null deviance: 2240.3  on 1110  degrees of freedom
Residual deviance: 2180.4  on 1098  degrees of freedom
AIC: 5621.5

Number of Fisher Scoring iterations: 10
```

*Figura 44. Significancia estadística de los parámetros del modelo de severidad IoT 1. Elaboración propia*

```
> Anova(modelo.sev_IoT_1_gamma, type=3)
Analysis of Deviance Table (Type III tests)

Response: AMT_Claim.x/1e+06
                       LR Chisq Df Pr(>Chisq)
Insured.age              0.5046  1    0.47749
Car.age                  5.0139  1    0.02514 *
Insured.sex              1.0444  1    0.30681
Marital                  1.0332  1    0.30940
Years.noclaims.y         0.0828  1    0.77353
Pct.drive.wkday          0.5476  1    0.45928
Accel.09miles            0.0530  1    0.81799
Accel.11miles            0.0918  1    0.76189
Brake.09miles            3.4675  1    0.06258 .
Brake.11miles            2.1343  1    0.14403
Left.turn.intensity10    0.0017  1    0.96687
Right.turn.intensity10   0.1719  1    0.67840
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 45 Tabla ANOVA del modelo de severidad IoT1*

- Modelo IoT 2:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.484e+01  7.238e-01  20.501   <2e-16 ***
Pct.drive.wkday       6.490e-01  9.574e-01   0.678   0.4980
Accel.09miles         4.902e-03  3.549e-02   0.138   0.8902
Accel.11miles         1.807e-02  7.815e-02   0.231   0.8172
Brake.09miles         6.998e-02  3.497e-02   2.001   0.0456 *
Brake.11miles        -1.317e-01  8.604e-02  -1.531   0.1261
Left.turn.intensity10 3.701e-07  3.114e-06   0.119   0.9054
Right.turn.intensity10 1.573e-06  3.771e-06   0.417   0.6767
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.800743)

    Null deviance: 2240.3  on 1110  degrees of freedom
Residual deviance: 2210.3  on 1103  degrees of freedom
AIC: 36329

Number of Fisher Scoring iterations: 10
```

*Figura 46. Significancia estadística de los parámetros del modelo de severidad IoT 2. Elaboración propia*

```
> Anova(modelo.sev_IoT_2_gamma, type=3)
Analysis of Deviance Table (Type III tests)

Response: AMT_Claim.x
                      LR Chisq Df Pr(>Chisq)
Pct.drive.wkday         0.3621  1    0.54734
Accel.09miles           0.0236  1    0.87793
Accel.11miles           0.0546  1    0.81528
Brake.09miles           4.1344  1    0.04202 *
Brake.11miles           2.3307  1    0.12685
Left.turn.intensity10   0.0157  1    0.90017
Right.turn.intensity10  0.2204  1    0.63872
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figura 47. Tabla ANOVA del modelo de severidad IoT2*

- Modelo IoT 3:
  - Step AIC dirección “Both”

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.61012    0.09712 160.729   <2e-16 ***
Car.age     -0.04516    0.01915  -2.358   0.0186 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.86314)

    Null deviance: 2250.5  on 1110  degrees of freedom
Residual deviance: 2227.2  on 1109  degrees of freedom
AIC: 36327

Number of Fisher Scoring iterations: 7
```

*Figura 48. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Both". Elaboración propia*

```
> Anova(modelo.sev_IoT_3_gamma_both, type=3)
Analysis of Deviance Table (Type III tests)

Response: AMT_Claim.x
         LR Chisq Df Pr(>Chisq)
Car.age    6.0395  1    0.01399 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 49  Tabla ANOVA del modelo de severidad IoT3 Both*

- Step AIC dirección “Backward”

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.59826    0.12431  12.857   <2e-16 ***
Car.age        -0.04086    0.01936  -2.110   0.0351 *
Insured.sexMale 0.12973    0.11910   1.089   0.2763
Brake.09miles   0.05735    0.02336   2.454   0.0143 *
Brake.11miles  -0.08837    0.04434  -1.993   0.0465 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.857526)

    Null deviance: 2240.3  on 1110  degrees of freedom
Residual deviance: 2192.1  on 1106  degrees of freedom
AIC: 5613

Number of Fisher Scoring iterations: 9
```

*Figura 50. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Backward". Elaboración propia*

```
> Anova(modelo.sev_IoT_3_gamma_bw, type=3)
Analysis of Deviance Table (Type III tests)

Response: AMT_Claim.x/1e+06
              LR Chisq Df Pr(>Chisq)
Car.age         4.7915  1    0.02860 *
Insured.sex     1.1793  1    0.27751
Brake.09miles   5.6307  1    0.01765 *
Brake.11miles   3.7744  1    0.05204 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 51  Tabla ANOVA del modelo de severidad IoT3 Backward*

- Step AIC dirección “Forward”

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.585124   0.111232 131.124  < 2e-16 ***
Car.age     -0.036687   0.009996  -3.670 0.000255 ***
Insured.age  0.006139   0.002307   2.661 0.007914 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.9245974)

    Null deviance: 1164.1  on 993  degrees of freedom
Residual deviance: 1144.9  on 991  degrees of freedom
AIC: 31246

Number of Fisher Scoring iterations: 6
```

*Figura 52. Significancia estadística de los parámetros del modelo de severidad IoT 3 usando StepAIC en dirección "Forward". Elaboración propia*

```
> Anova(modelo.sev_IoT_3_gamma_fw, type=3)
Analysis of Deviance Table (Type III tests)

Response: AMT_Claim.x
        LR Chisq Df Pr(>Chisq)
Car.age   6.0395  1    0.01399 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 53  Tabla ANOVA del modelo de severidad IoT3 Forward*

## Anexo 4 Resultados de los modelos de Tweedie

- Modelo base:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.6383966  0.0177993 148.230  < 2e-16 ***
Insured.age       0.0010370  0.0004718   2.198 0.028164 *
Car.age          -0.0037380  0.0011147  -3.353 0.000826 ***
Insured.sexMale   0.0086658  0.0069259   1.251 0.211126
MaritalSingle     0.0047198  0.0080188   0.589 0.556256
Years.noclaims.y -0.0005943  0.0004293  -1.384 0.166550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 0.01437796)

    Null deviance: 17.898  on 1110  degrees of freedom
Residual deviance: 17.644  on 1105  degrees of freedom
AIC: NA
```

*Figura 54. Significancia estadística de los parámetros del modelo Tweedie base. Elaboración propia*

```
> Anova(modTwee_trad, type=3)
Analysis of Deviance Table (Type III tests)

Response: log(AMT_Claim.x)
                 LR Chisq Df Pr(>Chisq)
Insured.age        4.8098  1  0.0282991 *
Car.age           11.3584  1  0.0007511 ***
Insured.sex        1.5643  1  0.2110371
Marital            0.3464  1  0.5561706
Years.noclaims.y   1.9077  1  0.1672146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 55 Tabla ANOVA del modelo Tweedie base*

- Modelo IoT 1:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.625e+00  4.527e-02  57.987  < 2e-16 ***
Insured.age          1.037e-03  4.769e-04   2.175 0.029872 *
Car.age             -3.814e-03  1.138e-03  -3.353 0.000828 ***
Insured.sexMale      9.466e-03  6.972e-03   1.358 0.174820
MaritalSingle        6.068e-03  8.052e-03   0.754 0.451220
Years.noclaims.y    -5.154e-04  4.421e-04  -1.166 0.244001
Pct.drive.wkday      1.203e-02  5.695e-02   0.211 0.832820
Accel.09miles       -3.795e-03  2.081e-03  -1.823 0.068551 .
Accel.11miles        1.298e-02  4.578e-03   2.836 0.004656 **
Brake.09miles        5.976e-03  2.071e-03   2.886 0.003978 **
Brake.11miles       -1.625e-02  5.041e-03  -3.224 0.001304 **
Left.turn.intensity10  9.779e-08  1.821e-07   0.537 0.591378
Right.turn.intensity10 2.259e-07  2.199e-07   1.027 0.304572
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 0.01434817)

    Null deviance: 17.898  on 1110  degrees of freedom
Residual deviance: 17.464  on 1098  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

*Figura 56. Significancia estadística de los parámetros del modelo Tweedie IoT 1. Elaboración propia*

```
> Anova(modTwee_IoT_1, type=3)
Analysis of Deviance Table (Type III tests)

Response: log(AMT_Claim.x)
                     LR Chisq Df Pr(>Chisq)
Insured.age            4.7061  1  0.0300563 *
Car.age               11.3516  1  0.0007539 ***
Insured.sex            1.8418  1  0.1747366
Marital                0.5679  1  0.4510800
Years.noclaims.y       1.3510  1  0.2451057
Pct.drive.wkday        0.0442  1  0.8335155
Accel.09miles          3.4075  1  0.0649016 .
Accel.11miles          8.0760  1  0.0044856 **
Brake.09miles          8.3118  1  0.0039389 **
Brake.11miles         10.3345  1  0.0013057 **
Left.turn.intensity10  0.2941  1  0.5875850
Right.turn.intensity10 1.0880  1  0.2969071
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figura 57 Tabla ANOVA del modelo Tweedie IoT1*

- Modelo IoT 2:

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.648e+00  4.240e-02  62.461  < 2e-16 ***
Pct.drive.wkday      1.675e-02  5.609e-02   0.299  0.76523
Accel.09miles       -3.605e-03  2.079e-03  -1.734  0.08314 .
Accel.11miles        1.166e-02  4.577e-03   2.547  0.01101 *
Brake.09miles        6.117e-03  2.048e-03   2.986  0.00289 **
Brake.11miles       -1.595e-02  5.040e-03  -3.165  0.00159 **
Left.turn.intensity10  1.411e-07  1.823e-07   0.774  0.43892
Right.turn.intensity10 2.187e-07  2.206e-07   0.991  0.32176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 0.01451543)

    Null deviance: 17.898  on 1110  degrees of freedom
Residual deviance: 17.725  on 1103  degrees of freedom
AIC: NA
```

*Figura 58. Significancia estadística de los parámetros del modelo Tweedie IoT 2. Elaboración propia*

```
> Anova(modTwee_IoT_2, type=3)
Analysis of Deviance Table (Type III tests)

Response: log(AMT_Claim.x)
                     LR Chisq Df Pr(>Chisq)
Pct.drive.wkday        0.0883  1   0.766352
Accel.09miles          3.0808  1   0.079220 .
Accel.11miles          6.5021  1   0.010775 *
Brake.09miles          8.8947  1   0.002860 **
Brake.11miles          9.9666  1   0.001594 **
Left.turn.intensity10  0.6144  1   0.433120
Right.turn.intensity10 1.0130  1   0.314196
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 59 Tabla ANOVA del modelo Tweedie IoT2*

- Modelo IoT 3: Dirección backward

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.6531868  0.0129228 205.310  < 2e-16 ***
Insured.age    0.0005206  0.0002606   1.998 0.045948 *
Car.age       -0.0038815  0.0011311  -3.431 0.000622 ***
Accel.09miles -0.0036712  0.0020654  -1.777 0.075764 .
Accel.11miles  0.0128369  0.0045534   2.819 0.004900 **
Brake.09miles  0.0062209  0.0020395   3.050 0.002341 **
Brake.11miles -0.0164280  0.0050083  -3.280 0.001070 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 0.01433261)

    Null deviance: 17.898  on 1110  degrees of freedom
Residual deviance: 17.536  on 1104  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

*Figura 60. Significancia estadística de los parámetros del modelo Tweedie usando StepAIC en dirección Backward. Elaboración propia*

```
> Anova(modTwee_IoT_3, type=3)
Analysis of Deviance Table (Type III tests)

Response: log(AMT_Claim.x)
               LR Chisq Df Pr(>Chisq)
Insured.age      3.9838  1  0.0459400 *
Car.age         11.8927  1  0.0005635 ***
Accel.09miles    3.2396  1  0.0718791 .
Accel.11miles    7.9680  1  0.0047613 **
Brake.09miles    9.2687  1  0.0023310 **
Brake.11miles   10.6824  1  0.0010816 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figura 61 Tabla ANOVA del modelo Tweedie IoT3*