UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA NEGOCIOS INTERNACIONALES

Título:

Modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación del consumidor

AUTORES:

Álava Llusca, Isabella Dominique

Gordon Sánchez, Flavio Paúl

Trabajo de Integración curricular previo a la obtención de título de

LICENCIADO EN NEGOCIOS INTERNACIONALES

TUTOR:

Ing. Carrera Buri, Félix Miguel, Mgs.

Guayaquil, Ecuador

Febrero 07, 2025

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA NEGOCIOS INTERNACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de integración curricular fue realizado en su totalidad por **Álava Llusca, Isabella Dominique y Gordon Sánchez, Flavio Paúl,** como requerimiento para la obtención del título de **Licenciados en Negocios Internacionales**.

TUTOR (A)

f. _____

Ing. Carrera Buri, Félix Miguel, Mgs.

DIRECTORA DE LA CARRERA

f. _____

Ing. Hurtado Cevallos, Gabriela Elizabeth, Mgs.

Guayaquil, a los 07 del mes de febrero del año 2025

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA NEGOCIOS INTERNACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Nosotros, **Álava Llusca, Isabella Dominique y Gordon Sánchez, Flavio Paúl**

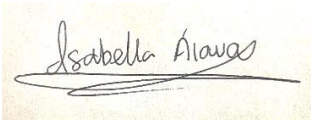
DECLARAMOS QUE:

El Trabajo de Integración Curricular, **Modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación del consumidor,** previo a la obtención del título de **Licenciados en Negocios Internacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

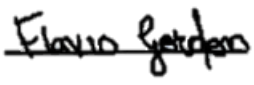
Guayaquil, a los 07 del mes de febrero del año 2025

LOS AUTORES:



f. _____

Álava Llusca, Isabella Dominique



f. _____

Gordon Sánchez, Flavio Paúl

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA NEGOCIOS INTERNACIONALES

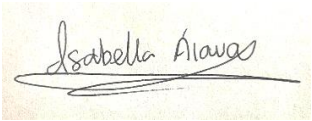
AUTORIZACIÓN

Nosotros, **Álava Llusca, Isabella Dominique y Gordon Sánchez, Flavio Paúl**

Autorizamos a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Integración Curricular, **Modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación del consumidor**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

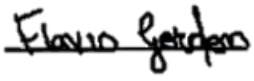
Guayaquil, a los 07 del mes de febrero del año 2025

LOS AUTORES:



f. _____

Álava Llusca, Isabella Dominique



f. _____

Gordon Sánchez, Flavio Paúl

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA NEGOCIOS INTERNACIONALES

REPORTE COMPILATIO

CERTIFICADO DE ANÁLISIS
magister

Tesis Final Flavio Paúl Gordon Sánchez-Isabella Dominique Álava Llusca

9% Textos sospechosos

100% Similitudes (ignorado)
2% similitudes entre comillas
9% entre las fuentes mencionadas
4% Idiomas no reconocidos
5% Textos potencialmente generados por la IA

Nombre del documento: Tesis Final Flavio Paúl Gordon Sánchez-Isabella Dominique Álava Llusca.docx
ID del documento: 42e89d5d72b22f378ff88ef07b086e38e4a61a86
Tamaño del documento original: 3,18 MB
Autores: []

Depositante: Felix Miguel Carrera Buri
Fecha de depósito: 5/2/2025
Tipo de carga: interface
fecha de fin de análisis: 5/2/2025

Número de palabras: 18.366
Número de caracteres: 124.878

Ubicación de las similitudes en el documento:

Ing. Carrera Buri, Félix Miguel, Mgs.

Agradecimiento

Quiero agradecer a mi mamá y a mi abuelito Papi Pepe por haberme otorgado la oportunidad de estudiar y formarme como una profesional.

A mi compañero de tesis, Flavio Paúl Gordon Sánchez, quien ha estado conmigo en los mejores y peores momentos. Gracias por ser mi apoyo incondicional.

Esta etapa ha sido una mezcla de emociones que recordaré con mucho cariño, como mi primera lloradera por sacar baja nota en una tutoría de estadística, mis nervios en todas las clases de francés porque solo sabía decir "oui", por aventurarme en irme de intercambio y siempre ir a comer a San Pedro después de clases.

Sobre todo, estoy profundamente agradecida por las personas que conocí en el camino y que hicieron de esta etapa increíble. Mi Grupi: Juan Martín, Sarita y Dani. Gracias por no dejarme sola, por acolitarme en todo, por verle el lado positivo a la situación cuando teníamos mil pendientes encima. En serio me siento tan afortunada de haber coincidido con ustedes que sin dudarlo repetiría esta experiencia.

- *Álava Llusca, Isabella Dominique*

Dedicatoria

Este trabajo se lo dedico a mi mami, Mónica Guadalupe Llusca Toledo, quien ha creído en mí cuando yo no lo hacía, por apoyarme en todas mis decisiones y brindarme palabras de aliento cuando más las necesitaba. Espero que te sientas orgullosa de mí, así como yo me siento de ser tu hija.

A mi abuelito Papi Pepe, quien me permitió estudiar en la universidad y desde siempre me ha dicho que puedo con todo lo que me proponga. A mi tía Kary, quien después de mucho pensarlo, me ayudó a escoger la carrera. Los llevo en mi corazón y sé que están acompañándome desde arriba en cada paso que dé.

- *Álava Llusca, Isabella Dominique*

ÍNDICE

ÍNDICE FIGURAS

ÍNDICE TABLAS

ÍNDICE GRÁFICOS

Resumen

El presente trabajo investigativo utiliza los conceptos y algoritmos de Machine Learning, Regresión Lineal Múltiple y K-Means, para mejorar la precisión en el cálculo de las primas vehiculares y la segmentación de sus respectivos clientes. Basado en esto, se desarrolló un modelo de Aprendizaje Automatizado, que predice a través de variables como el sexo, valor vehículo, valor de prima actual, entre otras, cuanto sería la prima por pagar de cada cliente al momento de asegurar el vehículo. Además, con estas mismas variables el modelo también segmenta a los clientes dependiendo de sus características, dando así una mayor personalización de las primas vehiculares y determinado las categorías de posibilidad a pagar en los clientes, permitiendo así que futuros clientes que puedan tener características similares, se pueda tener una idea clara de cuanto deben de pagar por el valor de la prima vehicular.

Palabras clave: Industria seguros vehiculares, machine learning, prima vehicular.

Abstract

This research uses the concepts and algorithms of Machine Learning, Multiple Linear Regression and K-means, to improve the accuracy in the calculation of vehicle premiums and the segmentation of their respective clients. Based on this, an Automated Learning model was developed, which predicts through variables such as sex, vehicle value, current premium value, among others, how much the premium would be to be paid by each client at the time of insuring the vehicle. In addition, with these same variables the model also segments clients depending on their characteristics, thus giving greater personalization of vehicle premiums, and determining the categories of possibility to pay in clients, thus allowing future clients to have similar characteristics. one can have a clear idea of how much they should pay for the value of the vehicle premium.

Keywords: Vehicle insurance industry, machine learning, vehicle premium.

Introducción

El seguro vehicular tiene sus orígenes a finales del siglo XIX, donde los vehículos eran un producto novedoso, pero con muchas deficiencias a comparación de lo que conocemos hoy en día. Por consiguiente, se crea este mecanismo como respuesta al creciente uso de automóviles y la necesidad de proteger a conductores y terceros ante posibles accidentes. Sorprendentemente, el primer país en crear de la prima vehicular no sería la *Locomotora de Europa*, sino Estados Unidos. Otros autores han afirmado lo siguiente:

> Según la Oficina del Censo de EE. UU., la primera póliza de seguro de automóvil vendida en los EE. UU. fue vendida por Travelers en 1898. El Dr. Truman Martin de Buffalo, Nueva York, quien utilizaba su vehículo para visitas domiciliarias, compró la póliza y compró $5000 en cobertura de responsabilidad por $12,25, que es el equivalente a más de $347,89 dólares actuales. (Fernández, 2016, párr. 1)

Por otra parte, en Europa, los tribunales estuvieron a favor de la creación de seguros vehiculares, donde se pactó los siguientes fundamentos: “Los seguros solo podían cubrir las responsabilidades civiles de los conductores, pero en ningún caso sus responsabilidades penales, por ello no procedía prohibir estos seguros, dado que no era cierto que indujeran a la comisión de un delito por parte del asegurado” (La Vanguardia, 2019, párr. 6).

Paralelamente, se creó una ley titulada *Road Traffic Act 1930* (Ley de Tráfico Vial de 1930), donde la misma establecía lo siguiente:

> Una ley para establecer disposiciones para la regulación del tráfico en carreteras y de vehículos de motor y de otras maneras con respecto a carreteras y vehículos en ellas, para establecer disposiciones para la protección de terceros contra riesgos que surjan

> del uso de vehículos de motor y en conexión con dicha protección para enmendar la Ley de Compañías de Seguros de 1909 para enmendar la ley con respecto a los poderes de las autoridades locales para proporcionar vehículos de servicio público, y para otros fines relacionados con los asuntos antes mencionados. (Gobierno de Reino Unido, 1930, p.1)

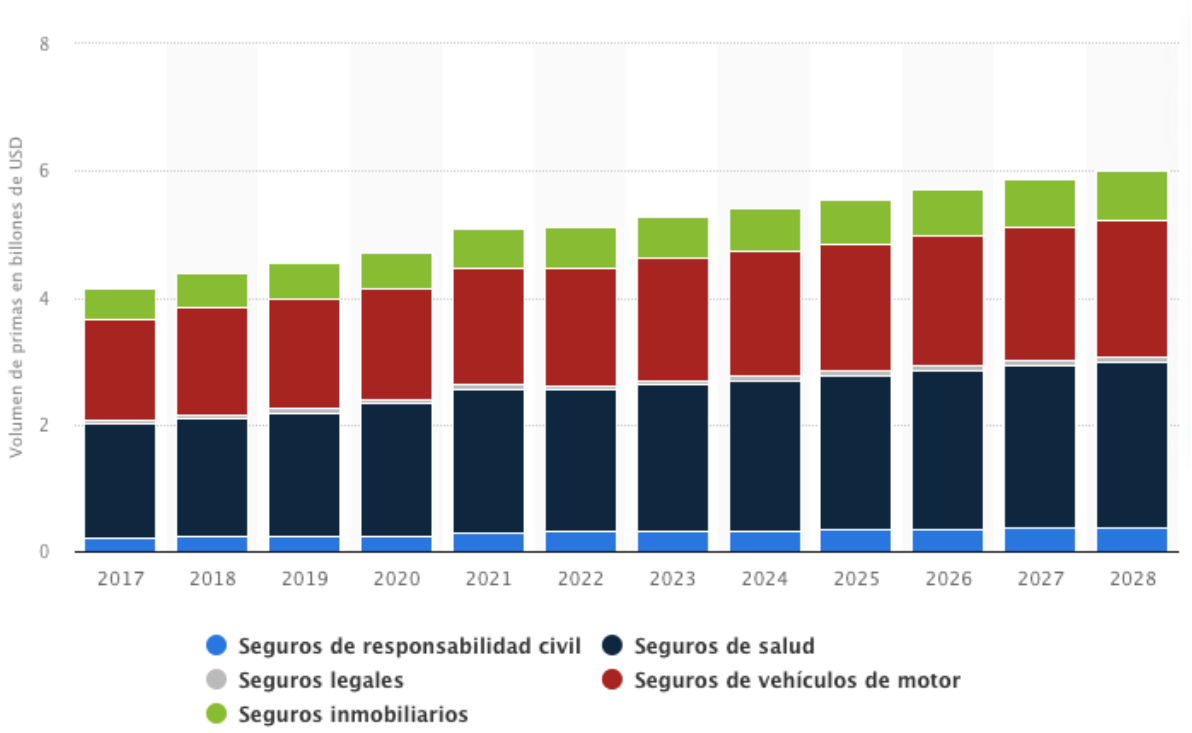
Sin duda alguna, este reglamento fue un ejemplar para el mundo, donde todos pudieron guiarse de esta medida y la implementaron con ciertas adaptaciones al contexto de cada una de sus naciones, por mencionar, la extensión de cobertura de daños, la prestación de servicios y otros temas relacionados al negocio.

Ahora bien, es pertinente destacar las diferencias entre los seguros. De acuerdo con la publicación del grupo financiero BBVA (2024) señala que:

> En los seguros de vida se suele ofrecer una compensación económica por la muerte natural o accidental (incluso ambas) del titular del seguro. También hay seguros de vida que incluyen compensaciones por enfermedades críticas o incapacidades. Por su parte, los seguros de no vida ofrecen otro tipo de coberturas destinadas a salvaguardar la integridad de bienes materiales (coche, hogar, etc.) o la salud del asegurado (es el caso de los seguros médicos). (p.1)

Con esto aclarado, los seguros de no vida toman un rol importante dentro de la economía del mercado mundial, así sean seguros legales, inmobiliarios, vehiculares, de salud y de responsabilidad civil. Es una industria que atribuye billones de dólares cada año, incluso hasta ahora refleja un aporte de aproximadamente 5,41 billones de dólares estadounidenses. En el gráfico a continuación, se evidencia la participación de cada uno de los mencionados.

Figura 1

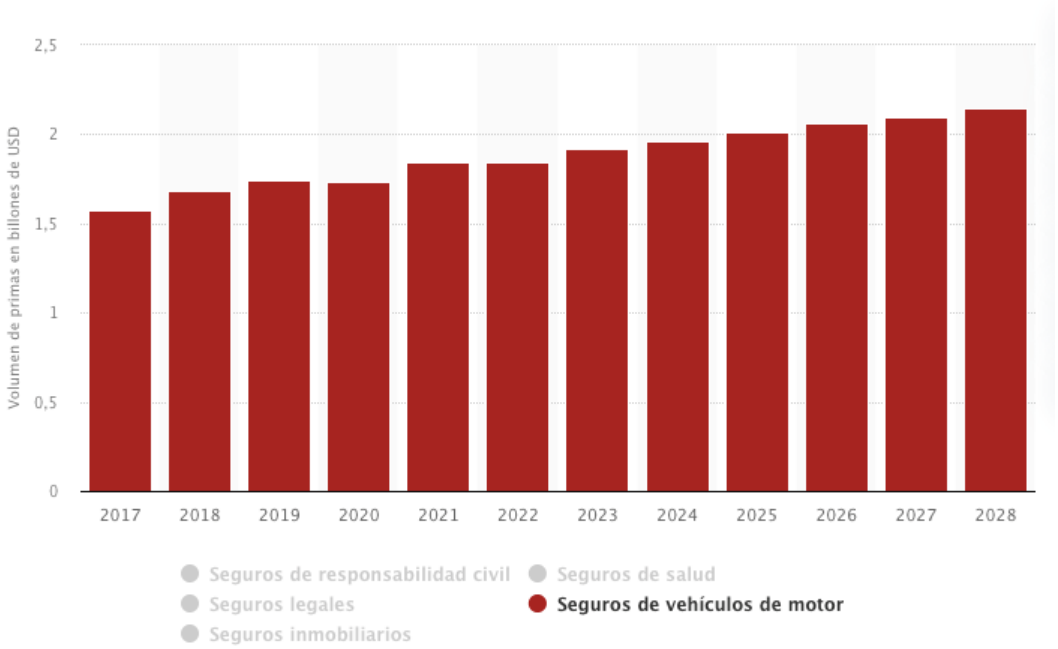


Ingresos por la emisión de primas del mercado mundial de seguros de no vida desde 2017 hasta 2028, por segmento. Statista, 2024.

Ahora bien, este trabajo se centra en los seguros de automóviles, es por ello que, en el gráfico de abajo, se filtró por "Seguros de vehículos de motor" y claramente se puede observar que es uno de los principales contribuyentes, en conjunto con el seguro de salud.

Los seguros vehiculares se han mantenido en un rango promedio, es decir, desde el año 2017 hasta lo que va del año 2024, la industria ha reflejado los siguientes valores en billones de dólares estadounidenses: 1,57; 1,58; 1,74; 1,73; 1, 84; 1,84; 1, 92 y 1,96 respectivamente.

Figura 2



Ingresos por la emisión de primas del mercado mundial de seguros de no vida desde 2017 hasta 2028, por segmento. Statista, 2024.

Incluso, se puede observar lo que se pronostica para los años siguientes, en el 2028 se espera que en general, la contribución de los seguros de no vida llegue a 6 billones de dólares estadounidenses, los seguros vehiculares aportarían con 2,15 millones de dólares.

Enfocando el tema a un nivel meso, América del Sur no se queda atrás, alrededor de 14 países hacen uso obligatorio de las primas. De acuerdo con un artículo de La República (2022), menciona lo siguiente:

> Los seguros más caros de este último tipo se encuentran en Brasil, donde su costo en dólares - que puede variar según la tasa de cambio - es de US$548 para automóviles y de US$221 para motos. En el segundo lugar le sigue Perú, con un valor de US$140 y US$245, respectivamente; seguido por Colombia, con US$127 y US$126; Panamá,

> con US$109 y US$40; y Uruguay, con US$70,6 y US$28. En contraste, aquellas zonas en donde el conocido localmente como Soat es más económico es Chile, con un costo aproximado de US$4,03 para automóviles y de US$42,5 para motos. Allí también destacan Bolivia, con US$13 y US$29, respectivamente; además de Ecuador, donde su valor asciende a US$37 y US$26, para cada tipo de automotor. (p.1)

Agregando a lo anterior, la pandemia fue un choque fuerte para el Ecuador. Bajo las palabras del economista Santiago Cobo (2022), menciona que:

> En el año 2020 el seguro de vehículos reflejó una producción de USD 341,5 millones, es decir 60 millones menos que el año 2019. Los ecuatorianos al sufrir la pandemia restaron prioridad a estas pólizas, dato que se comprueba con esta caída del casi 15%. En cuanto al resultado técnico del ramo, las 22 aseguradoras autorizadas en el país para operar el mismo, generaron resultados positivos en el año 2020 ya que el costo total de siniestros cayó en 27,4% originando cifras muy positivas en resultado técnico. (p.1)

En Ecuador las aseguradoras se encuentran en medio de un auge de seguros vehiculares. Dentro del territorio nacional, se pueden encontrar un sinnúmero de compañías cuyos servicios son brindar seguros vehiculares, entre ellas se destacan: Chubb Seguros Ecuador S.A., Generali Group, Zurich Seguros Ecuador, Sweaden Seguros, Seguros Equinoccial, AIG Metropolitana, MAPFRE Ecuador; todas controladas bajo la responsabilidad de la Superintendencia de Compañías Valores y Seguros.

Es evidente que las diferencias en los costos de los seguros dependen de ciertos aspectos como las condiciones políticas o económicas del país, la siniestralidad, entre otros;

sin embargo, existen factores indispensables para este cálculo, como lo es el perfil del conductor. Bajo este contexto, se puede observar la creciente tendencia hacia una personalización de seguros mediante el uso del aprendizaje automatizado.

"Un informe reciente afirma que, con una tasa de crecimiento anual compuesta (CAGR) del 32,5 %, es probable que la IA y el aprendizaje automático en el sector de los seguros generen 45 740 millones de dólares en todo el mundo en 2031" (ProjectPro, 2024, p.1); esto demuestra que habrá una gran demanda por estas tecnologías y las empresas deben estar actualizadas para mantenerse competitivos dentro del mercado globalizado.

En síntesis, el aprendizaje automatizado es revolucionario para el sector de seguros vehiculares, ya que busca optimizar procesos, ahorrar tiempo, disminuir errores y ofrecer un mejor servicio a sus usuarios. Se debe invertir en esta clase de tecnologías fusionada con el cálculo de la prima vehicular y segmentación del cliente para brindar eficacia a los negocios envueltos en este ámbito.

Problemática

El sector de seguros vehiculares en Ecuador enfrenta desafíos significativos que afectan tanto a las compañías aseguradoras como a los consumidores. En primer lugar, las metodologías tradicionales utilizadas para calcular las primas de seguro capturan cierta complejidad y diversidad de los riesgos actuales y las necesidades de los clientes. Estas metodologías suelen basarse en variables como el precio del vehículo, la siniestralidad, sin considerar factores dinámicos y comportamientos específicos de los conductores (Superintendencia de Compañías, Valores y Seguros, 2020).

Como consecuencia, conductores de bajo riesgo probablemente estarían pagando una prima considerablemente alta, mientras que, los conductores de alto riesgo pagan menos de lo que deberían. Esto al corto y largo plazo, refleja una inequidad entre los usuarios, los cuales buscarán otras empresas que lo segmenten adecuadamente, bajo las características de cada uno.

De acuerdo con FasterCapital (s.f.) define las consecuencias de los conductores de alto riesgo, tales como:

> Ser un conductor de alto riesgo puede tener consecuencias importantes, incluidas primas de seguro más altas, opciones de cobertura limitadas y dificultades para obtener un seguro. Las compañías de seguros cobran primas más altas a los conductores de alto riesgo para compensar la mayor probabilidad de presentar reclamaciones. Como resultado, los conductores de alto riesgo pueden tener dificultades para pagar un seguro o verse obligados a conformarse con opciones de cobertura limitadas. Además, algunas compañías de seguros pueden negarse por completo a asegurar a los conductores de alto riesgo, dejándolos sin cobertura de seguro. (p.1)

Como se ha mencionado en la previamente, el aprendizaje automatizado y el mercado de seguros vehiculares son una combinación estratégica, por ello, surge la siguiente duda: ¿Qué ocurriría si no se implementa el ML para el cálculo de la prima vehicular y segmentación del consumidor?

Existen varias consecuencias a esta interrogante. La principal contrariedad sería un desacierto total por la tecnología existente hoy en día para optimizar el proceso de cálculo.

No se quiere decir que los modelos tradicionales son sean útiles, porque por algo se han mantenido hasta la actualidad; sin embargo, los mismos tienen unas limitantes como:

> Personalización limitada: los modelos tradicionales pueden tener dificultades para ofrecer tarifas de primas personalizadas y adaptadas a cada asegurado, lo que puede dar lugar a una fijación de precios menos precisa para perfiles de riesgo específicos. Evaluación estática de riesgos: la evaluación de riesgos en los modelos tradicionales puede ser relativamente estática y depender de actualizaciones periódicas en lugar de análisis de datos en tiempo real. Sensibilidad a sesgos históricos: el análisis de datos históricos puede introducir sesgos y limitaciones, ya que puede no captar por completo las tendencias emergentes o la dinámica cambiante. (Raghav, 2024, párr. 10-13)

Más que un aspecto negativo, debería ser visto como una oportunidad de mejora por las compañías ecuatorianas y en general a nivel de América Latina, ya que, en países como Estados Unidos, el uso de ML en varios sectores ya es una realidad que brinda muchos beneficios.

También, sin la aplicación del ML se obtendrían resultados en largos plazos de tiempo, es decir, tomaría más tiempo de lo habitual en entregar una cotización o una referencia. Y el tiempo en este tipo de industria es primordial, ya que cada segunda cuenta, siempre hay la posibilidad de que el cliente se vaya con la competencia. Siguiendo por esta idea, el trabajo humano tiende a ser manual, por ende, generar errores, los cuales representan pérdida de tiempo y de dinero.

“Al automatizar tareas repetitivas y laboriosas, el aprendizaje automático libera tiempo y recursos. Los esfuerzos pueden redirigirse a actividades de mayor valor añadido” (Salesforce, s.f., párr. 6). Esta afirmación una vez más demuestra que sin aprendizaje automatizado las empresas no pueden ser eficaces, y necesitan de la tecnología para prosperar, más aún, manejando un gran volumen de datos.

Adicionalmente, las empresas ecuatorianas deben estar pendientes de cada actualización y nuevo método que se desarrolle en la industria para mantenerse competitivo y sobre todo, relevante. Incluso, tomando en cuenta que empresas extranjeras se están posicionando en el país, con una gran inversión, tanto de capital como de tecnología, los negocios locales no se pueden arriesgar a perder participación en el mercado.

Igual de pertinente, es señalar que el no calcular una prima adecuadamente puede afectar a la retención de clientes. Esto tiene que ver con la segmentación de clientes, mientras más acertado se agrupe a un conjunto de personas debido a sus características y similitudes, más preciso serán las pólizas destinadas a cada uno de ellos; de esta manera, la empresa refleja un compromiso con sus usuarios y les brinda un sentimiento de confianza. Con respecto a lo redactado en la Adsalsa (2024), se menciona lo que se obtiene con respecto a la función de la IA en este campo:

> Garantiza un sistema de aprendizaje y mejora continua, por lo que la segmentación se convierte en un proceso dinámico que mejora con el tiempo. Permite un análisis en tiempo real, por lo que las empresas pueden ajustar sus estrategias y acciones de marketing con mayor agilidad (adelantándose a la competencia). Disponer de un alto nivel de escalabilidad, por lo que es ideal aplicarla a la segmentación de todo tipo de empresas y proyectos. (p.1)

En general, la inteligencia artificial no debería ser considerada como una amenaza, debería complementar al trabajo que ya realizan los funcionarios de una empresa. Es una tecnología tan prometedora, que incluso en un artículo de Harvard escrito por Siegel (2023) redacta que:

> ¿Por qué el ML es una tecnología tan prometedora para mejorar la experiencia del cliente? Es simple: puede predecir los comportamientos de los clientes. La predicción como capacidad es el *Holy Grail* para prever las necesidades de cada cliente y personalizar los productos y servicios en consecuencia. Desde la perspectiva del consumidor, cuando se evitan los escollos éticos del ML, la predicción puede ser el antídoto definitivo contra la sobrecarga de información a la que todos nos enfrentamos todos los días. (párr. 4)

En síntesis, es primordial que las empresas inviertan en tecnología, así como lo hacen para el departamento de marketing, recursos humanos, entre otros; el invertir en algo novedoso como lo es el ML, será una transformación total para bien, tanto para la empresa y su rentabilidad, como para la satisfacción y fidelidad del cliente.

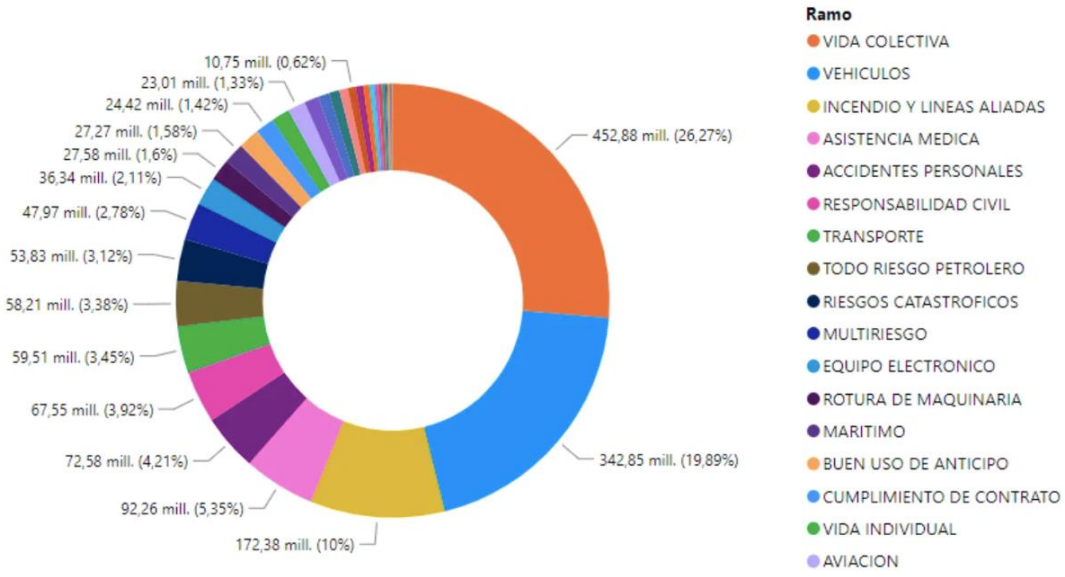
Justificación

La implementación de un modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación de clientes es crucial y necesaria para enfrentar las restricciones y limitaciones de las metodologías tradicionales con respecto al cálculo de la prima vehiculares, las cuales a lo largo de los años han venido perdiendo efectividad, generando así muchas dudas por parte del consumidor.

Utilizar la ciencia del machine learning para el cálculo de la prima vehicular y segmentación de clientes, no solo genera confianza en los consumidores de las pólizas de seguros, sino que también garantizan un mejor análisis individual de cada cliente, dando así una mayor precisión para el cálculo de la prima vehicular del mismo, y a su vez incluso le permite a la aseguradora medir el riesgo que conlleva firmar un contrato con un cliente en específico o no.

Es más, este proyecto es útil pertinente para el sector asegurador ecuatoriano, ya que su servicio se encuentra en auge. De acuerdo con los datos proporcionados por Actuaria (2022) reflejan lo siguiente:

Figura 3



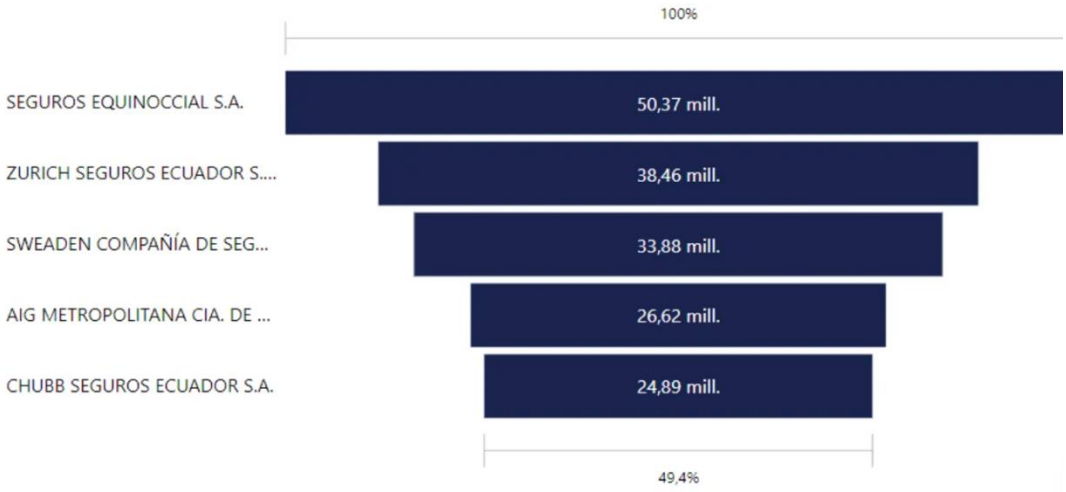
Ramos con mayor Prima Neta Emitida del último año. Actuaria, 2022.

Como se puede observar, en el año 2021, la mayoría de las primas emitidas fueron gracias al ramo de: vida colectiva (representando el 26,17% equivalente a 452,88 millones

de dólares), vehículos (representando el 19,89% equivalente a 342,85 millones de dólares) e incendio y líneas aliadas (representando el 10% equivalente a 172,38 millones de dólares).

De igual modo, los datos publicados por Actuaria demuestran la presencia de las aseguradoras vehiculares dentro del Ecuador.

Figura 4



Competitividad de principales empresas aseguradoras. Actuaria, 2022.

Este gráfico denota el nivel de competitividad entre las marcas mejor posicionadas dentro del país, una competencia que puede ser ganada dependiendo de cómo cada una de las empresas actúen ante los diferentes cambios globalizados.

Al incluir información sobre las variables sociodemográficas del cliente, el comportamiento y las condiciones de riesgo viales, las compañías de seguro pueden ofrecer unas pólizas más personalizadas que serán de mayor agrado para el cliente y el mismo se sentirá satisfecho con el servicio ofrecido por parte de la aseguradora. Fomentando así una mayor equidad y justicia con los clientes de pólizas y fortaleciendo así el vínculo de confianza entre el cliente y las aseguradoras vehiculares.

La segmentación detallada del cliente de seguro vehicular permite a las compañías de seguros la capacidad de desarrollar diferentes modelos de negocios y de productos que se adapten a las diferentes necesidades y opciones que buscan los clientes al momento de poder adquirir un seguro vehicular.

Los modelos de aprendizaje automatizado permiten detectar tendencias, patrones, gustos que poseen los clientes al momento que desean cotizar o adquirir algún producto del mercado asegurador, permitiendo así que las empresas aseguradoras puedan generar diferentes estrategias tanto de marketing como comerciales para los diferentes tipos de clientes que les ofrecen los resultados por medio del análisis de segmentación de clientes a través de machine learning, atendiendo así las demandas y necesidades particulares de cada segmento permite que a la empresa que adopte estas innovaciones pueda estar a la vanguardia del mercado, el cual es muy competitivo y se encuentra muy saturado de varias empresas ofreciendo este mismo servicio.

A su vez, otro beneficio fundamental que ofrece no solo la iniciativa como tal de este proyecto, si no en si la creación del modelo es el impacto en la eficiencia de las operaciones de las empresas aseguradoras.

La automatización de procesos no solo atraerá más clientes, sino que también permitirá un ahorra en las campañas que realice la empresa para adquirir los mismos, reduciendo así los gastos elevados en personal que no sea necesario, la equivocación humana y la agilidad en la toma de decisiones. Permitiendo así, que los recursos obtenidos de los ahorros de la eficiencia de las operaciones de la empresa puedan ser redirigidos a sectores estratégicos de la empresa, como el departamento de innovación, los cuales requieren un alto presupuesto y son de gran ayuda, crecimiento y desarrollo para la empresa.

Por otra parte, el previo reconocimiento y detección de fraude se potencia a través de este modelo automatizado. Estos algoritmos son capaces de una vez entrenados y nutridos con la información necesaria puedan detectar y alertar de patrones irregulares o dudosos, que pongan en peligro la economía de la compañía. Ya sea través de imágenes faciales o datos específicos como número de teléfono y correo electrónico, se puede evitar alguna suplantación de identidad que pueda generar al fraude económico y que perjudique tanto al cliente como a la compañía. Lo cual, genera confianza por parte de los clientes en la compañía, además que se garantizara la protección de sus datos personales.

La aplicación de un modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación del consumidor está completamente justificada por la extensión y alcance que llega a tener para desarrollar soluciones y abordar distintos temas dentro del mercado asegurador. Es una oportunidad para las organizaciones de innovar y optimizar sus recursos, mientras mejorar la calidad de la experiencia de sus usuarios.

Alcance

Actualmente, la industria de seguros vehiculares ecuatoriana enfrenta diversos desafíos considerables en cuanto a su eficiencia operativa. La falta de integración de tecnologías avanzadas, basadas en el análisis de datos, ha limitado el acoplamiento de estas empresas a los cambios y tendencias del mercado. Debido a esto, muchas de las empresas han venido perdiendo clientes, lo cual reduce cada vez más su presencia en el mercado local de seguros vehiculares. Por lo tanto, el incorporar las herramientas de Machine Learning representa una oportunidad muy crucial para mejorar la eficiencia operativa y sobre todo el correcto cobre o costo de las primas vehiculares a sus clientes. Permitiendo así, una gran aprobación de estos, y pudiendo llegar a más nichos de mercado, incluso a clientes que muy

probablemente nunca hubieran tenido un seguro vehicular debido a los costos elevados que han venido teniendo en los últimos años.

Por tanto, este proyecto pretende alcanzar a empresarios y trabajadores del sector de seguros vehiculares, que deseen mejorar y automatizar sus procesos de cálculos de primas vehiculares, y que pretendan generan una eficiencia en los mismos. Obteniendo así, primas vehiculares más justan, que permitirán generar mayor aceptación de los clientes actuales y generación de otros nuevos clientes interesados en obtener un seguro vehicular, lo que permitirá cumplir con los objetivos principales de la empresa, que uno de ellos es el constante crecimiento año a año. A su vez, los empresarios y personas involucradas en el mundo de los seguros vehiculares podrán utilizar este proyecto como guía para la correcta segmentación de sus clientes, lo que les permitirá tener en cuentas que sus productos o servicios estén acordes y se acoplen muy bien a las necesidades de cada tipo de cliente, en base a esta segmentación.

Así mismo, este proyecto busca aspirar a alcanzar a diversos estudiantes y jóvenes que manejen las diversas ramas de las ciencias de datos, que puedan utilizar esta tesis, para fomentar sus respectivos trabajos investigativos durante su proceso de aprendizaje o conocimiento. Lo estudiantes podrán utilizar este proyecto como parte inicial de futuras mejoras que se requieran en este campo de los seguros, garantizando así que futuramente existan modelos con mayor precisión y detalle que permitan aún más mejorar la forma en que calcula las primas vehiculares y se segmentas los clientes del sector vehicular de los seguros.

Objetivo General

Desarrollar un Modelo de Aprendizaje Automatizado para el Cálculo de la Prima Vehicular y la Segmentación de clientes.

Objetivos Específicos

- Investigar y proporcionar una sólida base teórica, conceptual y legal que respalde la creación del modelo de Aprendizaje Automatizado, aplicada al cálculo de la prima y segmentación de clientes del área de seguros vehiculares.
- Describir y elaborar la metodología de Regresión Lineal Múltiple y Kmeans para la aplicación de conceptos de Machine Learning.
- Evaluar el modelo y su eficiencia mediante la exactitud y precisión de los resultados de la predicción de la prima vehicular y la segmentación de los clientes.

Capítulo I

Marco Teórico

ETL

ETL (extracción, transformación, carga) es un proceso de integración de datos que se utiliza para combinar datos de varias fuentes en un conjunto de datos único y coherente para cargarlo en un almacén de datos, data lake u otro sistema de destino. **A** medida que las bases de datos ganaban popularidad en los años 70, se introdujo el ETL como proceso de integración y carga de datos para el cálculo y el análisis, que acabó convirtiéndose en el método principal para procesar datos en proyectos de almacenamiento de datos (Ibm, 2024).

ETL proporciona la base para los flujos de trabajo de análisis de datos y machine learning. A través de una serie de normas empresariales, ETL limpia y organiza los datos de una manera que satisface las necesidades específicas de inteligencia empresarial, como los informes mensuales, pero también puede abordar análisis más avanzados, que pueden mejorar los procesos de back-end o las experiencias de los usuarios finales (Ibm, 2024).

Figura 5

ETL. Área Tecnología, sf.

Machine Learning

El machine learning es la ciencia de desarrollo de algoritmos y modelos estadísticos que utilizan los sistemas de computación con el fin de llevar a cabo tareas sin instrucciones explícitas, basándose en patrones e inferencias. Los sistemas de computación utilizan algoritmos de machine learning para procesar grandes cantidades de datos históricos e identificar patrones de datos. Esto les permite generar resultados con mayor precisión a partir de un conjunto de datos de entrada. Por ejemplo, los científicos de datos pueden entrenar una aplicación médica para diagnosticar el cáncer con imágenes de rayos X a partir del almacenamiento de millones de imágenes escaneadas y diagnósticos correspondientes (AWS, 2024).

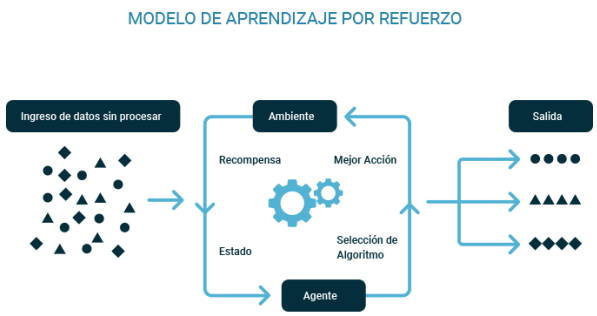
El machine learning permite que las empresas impulsen el crecimiento, generen nuevas fuentes de ingresos y resuelvan problemas complejos. Los datos son la fuerza que impulsa la toma de decisiones empresariales. Estos suelen tener diversos orígenes, como los comentarios de los clientes, los empleados y las finanzas. La investigación dedicada al machine learning automatiza y optimiza este proceso. Las empresas pueden obtener resultados más rápido con programas que analizan grandes volúmenes de datos a gran velocidad (AWS, 2024).

Aprendizaje No Supervisado

Los métodos no supervisados (unsupervised methods) son algoritmos que basan su proceso de entrenamiento en un juego de datos sin etiquetas o clases previamente definidas. Es decir, a priori no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. El aprendizaje no supervisado está dedicado a las tareas de agrupamiento, también llamadas

clustering o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos (Rueda, 2019).

Figura 6



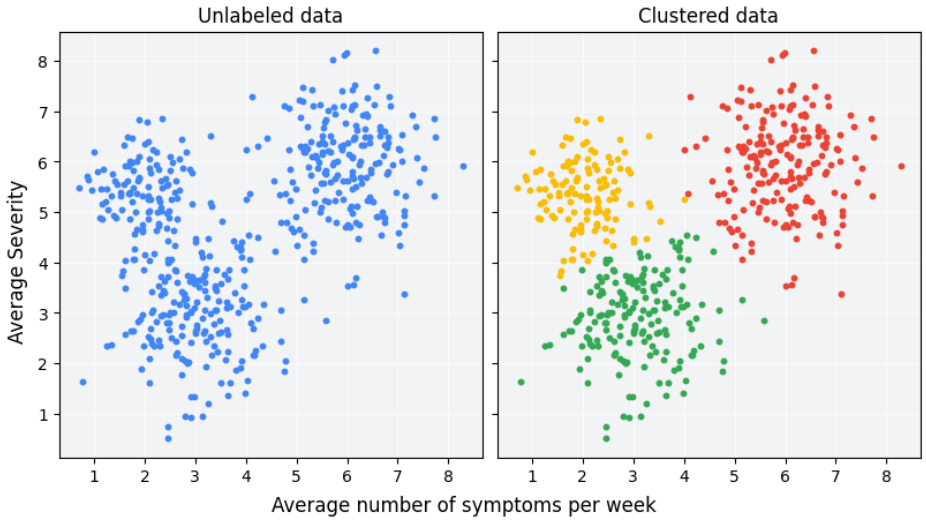
Tipos de aprendizaje automático. Medium, 2020.

Clustering

El clustering es un algoritmo de machine learning no supervisado que organiza y clasifica diferentes objetos, puntos de datos u observaciones en grupos o clusters basados en similitudes o patrones. Hay varias formas de usar el clustering en el machine learning, desde las exploraciones iniciales de un conjunto de datos hasta la monitorización de los procesos en curso (Ibm, 2024). Puede usarlo en el análisis de datos exploratorios con un nuevo conjunto de datos para comprender las tendencias, los patrones y los valores atípicos subyacentes. Como alternativa, puede tener un conjunto de datos más grande que deba dividirse en varios conjuntos de datos o reducirse mediante la reducción de dimensionalidad. En estos casos, la agrupación en clústeres puede ser un paso en el preprocesamiento. Los ejemplos de clústeres pueden incluir géneros musicales, diferentes grupos de usuarios, segmentos clave de una segmentación de mercado, tipos de tráfico de red en un clúster de

servidores, grupos de amigos en una red social o muchos otros tipos de categorías. El proceso de clustering puede usar solo una característica de los datos o puede usar todas las características presentes en los datos (Ibm, 2024).

Figura 7

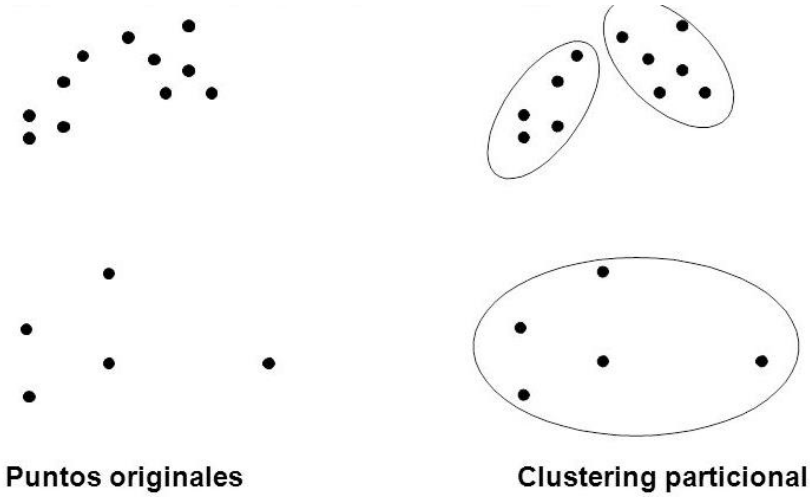


¿Qué es el agrupamiento en clústeres?, Google For Developers, s.f.

Clustering Particional

En el clustering particional el objetivo es obtener una partición de los objetos en grupos o clusters de tal forma que todos los objetos pertenezcan a alguno de los k clusters posibles y que por otra parte los clusters sean disjuntos (San Sebastián, s. f.).

Figura 8



Partitional clustering. ResearchGate, s.f.

Distancia Euclidiana

Este importante algoritmo conocido como algoritmo de distancia euclidiana es la aplicación de una fórmula matemática que permite medir la distancia en línea recta entre dos puntos en un espacio n-dimensional (GraphEverywhere, 2019).

$$\boldsymbol{d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}}$$

Dentro de las funcionalidades del algoritmo de distancia euclidiana se destaca su utilidad para determinar la similitud entre dos cosas o pares de datos. Partiendo de esto, puede utilizarse la similitud calculada a partir de este algoritmo como parte integral de sistemas de consulta de recomendación. Con ella podemos conseguir esquemas de datos que identifiquen elementos que tengan características similares, como una puntuación o una valoración para que el usuario de la información puede decidir. Un ejemplo de esto puede ser aplicado en un

sistema de recomendación de películas para que obtengamos sugerencias en base a las calificaciones recibidas (GraphEverywhere, 2019).

Distancia de Manhattan

La distancia de Manhattan, también conocida como distancia de taxi o distancia L1, es una métrica utilizada en varios campos como la estadística, análisis de los datos, y la ciencia de datos para medir la distancia entre dos puntos en un sistema basado en cuadrícula. Esta distancia se calcula sumando las diferencias absolutas de sus coordenadas cartesianas. El término "Manhattan" se deriva de la disposición en cuadrícula de las calles de Manhattan, en la ciudad de Nueva York, donde uno tendría que viajar a lo largo de las calles en lugar de en línea recta para llegar a un destino. Este concepto es particularmente útil en escenarios donde el movimiento está restringido a trayectorias horizontales y verticales, lo que lo convierte en una medida fundamental en la planificación urbana, la robótica y los gráficos por computadora (Learn Statistics Easily, 2024).

$$d(x_i - x_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

Distancia de Conseno

La similitud del coseno es una medida de similitud que se calcula entre dos vectores distintos de cero dentro del espacio interno del producto que mide el coseno del ángulo entre ellos. El coseno de 0° es 1, y es menor que uno para cualquier ángulo que se encuentre en el intervalo (0, π] radianes. Así que este caso se trata de un cálculo que da origen a un juicio de orientación y no de magnitud (GraphEverywhere, 2019).

$$Similitud\ Coseno = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2} \times \sqrt{\sum_{k=1}^{n} x_{jk}^2}}$$

Dentro de los datos más resaltantes que posee el cálculo de la similitud de coseno es que su nombre proviene del término «coseno de dirección» en este caso, los vectores unitarios cumplen con la condición de ser máximamente «similares» si son paralelos y en el caso de que sean máximamente diferentes, deben ser ortogonales o perpendiculares entre sí. Esto es similar a los valores que podemos encontrar en el coseno donde la unidad de valor máximo se encuentra cuando los segmentos tienen un ángulo cero y cero sin presentar alguna correlación, cuando los segmentos son totalmente perpendiculares (GraphEverywhere, 2019).

Método de Lloyd

El algoritmo de Lloyd es una técnica de agrupamiento que permite unificar n objetos en k clases. Al agrupar los objetos en distintos grupos o clases obtenemos un método de predicción que podemos aplicar tras encontrar un nuevo elemento. Así mismo podemos establecerlo en una de las clases conocidas, esperando que sus características sean similares a las del resto de elementos de dicha clase. Este algoritmo comienza dividiendo los elementos de entrada en las distintas clases iniciales, ya sea mediante el uso de datos heurísticos o mediante el azar, para proceder posteriormente con el cálculo del punto medio o centro de gravedad de cada conjunto (Colomé, 2012).

1. Inicialización de centroides $(\mu_1, \mu_2, \ldots, \mu_k)$.

2. Asignación de puntos al cluster más cercano:

$$C\_j=\{x_i : \|x_i-\mu_j\}^2 \leq \left|\left|x_i-\mu_l\right|\right|^2, \forall l \in \{1, \ldots., k\}\}$$

3. Actualización de centroides

$$: \mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

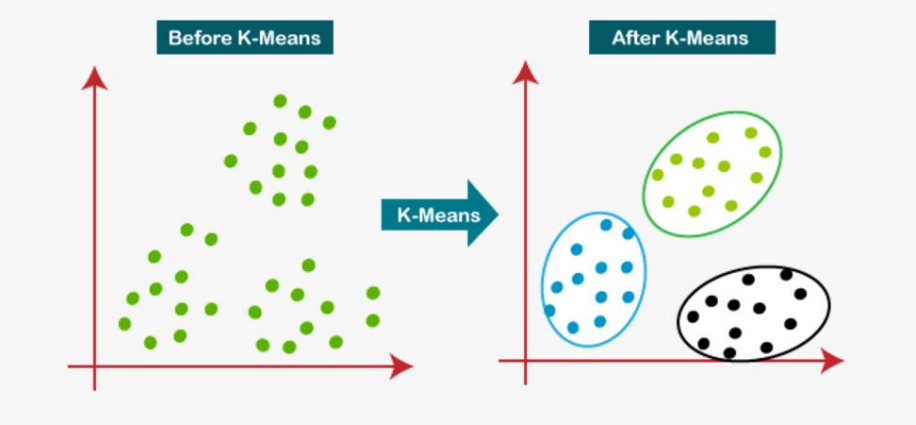
4. Repetición hasta la convergencia

De esta forma se construye una nueva partición, asociando cada elemento con el centro de gravedad más cercano y recalculando de nuevo los centroides. El algoritmo de Lloyd es pues un método de aprendizaje no supervisado debido a que nos permite obtener información sobre los objetos sin que aportemos de forma explícita un conocimiento previo sobre los mismos (Colomé, 2012).

K-means

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster (*RPubs - K Means*, s. f.).

Figura 9



K Means Clustering Algorithm. Prateek, 2022.

Método del codo (Elbow Method)

Según Jarroba, (2017):

> Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a N Clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide:

$$Inercia = \sum_{i=0}^{N} ||x_i - \mu||^2$$

Una vez obtenidos los valores de la inercia tras aplicar el K-means de 1 a N Clusters, representamos en una gráfica lineal la inercia respecto del número de Clusters. En esta gráfica se debería de apreciar un cambio brusco en la evolución de la inercia, teniendo la línea representada una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco en la inercia nos dirá el número óptimo de Clusters a seleccionar para ese data set; o, dicho de otra manera: el punto que representaría al codo del brazo será el número óptimo de Clusters para ese data set (Jarroba, 2017).

Coeficiente de Silueta

La Silhouette es una métrica que permite evaluar la calidad de los clústeres generados mediante algoritmos de clustering basados en la distancia euclídea. Como es el caso de k-means. Cuantificando la relación que existe entre la separación de los diferentes clústeres y la similitud entre los puntos de un mismo clúster en un valor que varía entre -1 y 1. Los valores cercanos a 1 indican la mejor separación de los clústeres y los cercanos a -1 la peor. Información que se puede utilizar para seleccionar el número óptimo de clústeres en k-means. Siendo una alternativa a otros métodos como el del codo (elbow method) (Rodríguez, 2023).

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i}$$

Aprendizaje Supervidado

Aprendizaje supervisado Es cuando entrenamos un algoritmo de Machine Learning dándole las preguntas (características) y las respuestas (etiquetas). Así en un futuro el algoritmo pueda hacer una predicción conociendo las características. En este tipo de aprendizaje hay dos algoritmos (entrenamientos): el de clasificación y el de regresión (Sandoval, 2018).

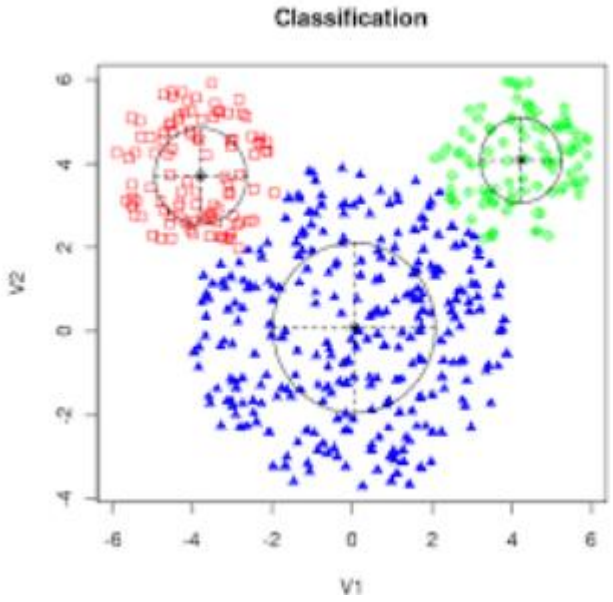
Algoritmo de Clasificación

De acuerdo con Sandoval (2018) menciona lo siguiente:

> Se espera que el algoritmo nos diga a qué grupo pertenece el elemento en estudio. El algoritmo encuentra patrones en los datos que le damos y los clasifica en grupos. Luego compara los nuevos datos y los ubica en uno de los grupos y es así como puede predecir de que se trata. La variable por predecir

es un conjunto de estados discretos o categóricos. Pueden ser: Binaria: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}, etc. Múltiple: Comprará {Producto1, Producto 2…}, etc. Ordenada: Riesgo {Bajo, Medio, Alto}, etc.

Figura 10

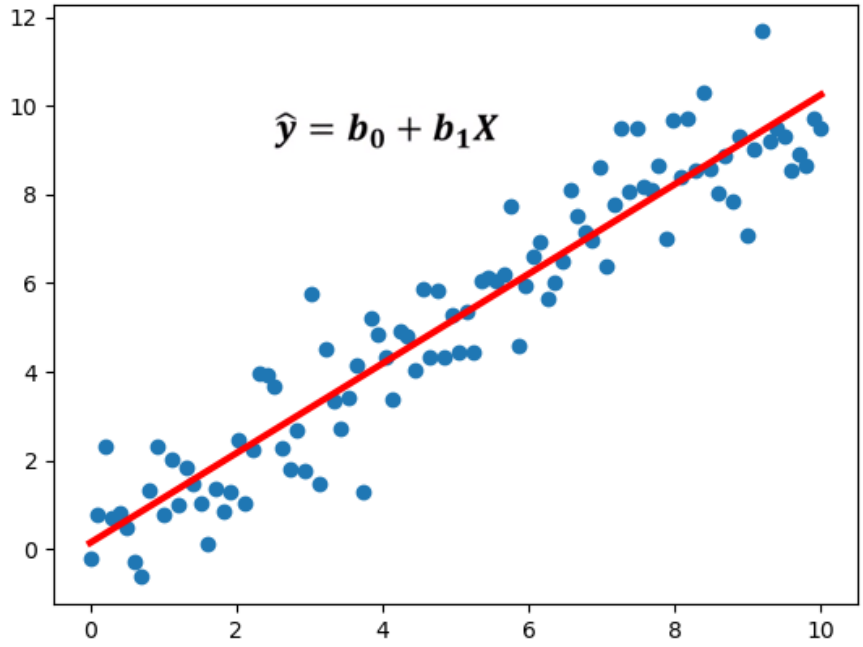


Aprendizaje supervisado y no supervisado. Blog de Francisco Sancho, s.f.

Algoritmo por Regresión

En el caso de los algoritmos de regresión, podemos decir que se trata de un subcampo del aprendizaje automático supervisado que tiene el fin de crear una metodología para relacionar un cierto número de características y una variable objetivo-continua. Algunos de los ejemplos más claros de los algoritmos de regresión son la estimación de cuánto tardará una persona en llegar a un destino, la predicción del tiempo que se mantendrá un empleado en una compañía (The Black Box Lab, 2022).

Figura 11



Modelo de regresión. Economipedia, 2022.

Correlación

La correlación es una asociación entre dos variables numéricas que evalúa la tendencia (aumento o disminución) en los datos. Cuando una variable nos proporciona información sobre otra variable, decimos que las dos variables están relacionadas. En cambio, cuando no hay correlación, el aumento o la disminución de una variable no nos dice nada sobre el comportamiento de otra variable (inbestMe, 2023).

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Análisis de Varianza (ANOVA)

El análisis de la varianza, o ANOVA, es un método de modelado lineal para evaluar la relación entre los campos. Para los controladores clave y para los conocimientos que están

relacionados con un número de gráficos, ANOVA prueba si el valor de destino medio varía entre las categorías de una entrada o combinaciones de categorías de dos entradas (*IBM Cognos Analytics 11.1.x*, s. f.).

Para probar si los medios son diferentes, una prueba ANOVA compara la varianza explicada (causada por los campos de entrada) con la varianza no explicada (provocada por el origen de error). Si la proporción de varianza explicada a la varianza no explicada es alta, los medios son estadísticamente diferentes (*IBM Cognos Analytics 11.1.x*, s. f.).

$$F = \frac{MSB}{MSW}$$

El estadístico F se compara con un valor crítico en una distribución F para determinar si las diferencias entre las medias son estadísticamente significativas.

Regresión Lineal Simple

En una regresión lineal, se trata de establecer una relación entre una variable independiente y su correspondiente variable dependiente. Esta relación se expresa como una línea recta. No es posible trazar una línea recta que pase por todos los puntos de un gráfico si estos se encuentran ordenados de manera caótica. Por lo tanto, sólo se determina la ubicación óptima de esta línea mediante una regresión lineal. Algunos puntos seguirán distanciados de la recta, pero esta distancia debe ser mínima. El cálculo de la distancia mínima de la recta a cada punto se denomina función de pérdida (Saavedra, 2023).

$$Y = \beta_0 + \beta_1 X + E$$

Regresión Polinómica

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + E}$$

La realidad es que la Regresión Polinomial extiende el modelo lineal al agregar predictores adicionales, que se obtienen al elevar cada uno de los predictores originales a una potencia. Por ejemplo, una regresión cúbica utiliza tres variables independientes, como predictores. Este enfoque proporciona una forma sencilla de proporcionar un ajuste no lineal a los datos. El método estándar para extender la Regresión Lineal a una relación no lineal entre las variables dependientes e independientes ha sido reemplazar el modelo lineal con una función polinomial (Gonzalez, 2022).

Regresión Lineal Múltiple

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (YY) se determina a partir de un conjunto de variables independientes llamadas predictores (X1X1, X2X2, X3X3…). Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella (esto último se debe que analizar con cautela para no malinterpretar causa-efecto) (Ciencia de Datos, s. f.).

Los modelos lineales múltiples siguen la siguiente ecuación:

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X_n + E}$$

Coeficiente de Determinación (R2R^2R2)

El coeficiente de determinación es la proporción de la varianza total de una variable explicada por una regresión. También conocido como R cuadrado, indica qué tan bien un modelo se ajusta a la variable que pretende explicar. El resultado del coeficiente de

determinación varía entre 0 y 1. Cuanto más cerca esté de 1, mejor se ajusta el modelo a la variable. Si está cerca de 0, el modelo es menos ajustado y fiable (López, 2024).

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y}_i)^2}$$

Error Medio Absoluto (MAE)

El error absoluto medio (MAE) es una métrica ampliamente utilizada en estadística y análisis de los datos que cuantifica la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Se define como el promedio de las diferencias absolutas entre los valores predichos y los valores reales. MAE proporciona una interpretación sencilla de la precisión de las predicciones, lo que la convierte en una herramienta esencial para los científicos de datos y los analistas a la hora de evaluar el rendimiento de los modelos de regresión. Al centrarse únicamente en el tamaño de los errores, MAE ayuda a identificar qué tan lejos están las predicciones de los resultados reales, lo que permite una comprensión más clara del rendimiento del modelo (Learn Statistics Easily, 2024).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Error Cuadrático Medio (MSE)

Error de raíz cuadrada media (RMSE) es la desviación estándar de los valores residuales (errores de predicción). Los valores residuales son una medida de la distancia de los puntos de datos de la línea de regresión; RMSE es una medida de cuál es el nivel de dispersión de estos valores residuales. En otras palabras, le indica el nivel de concentración

de los datos en la línea de mejor ajuste (*Oracle® Fusion Cloud EPM Trabajo Con Planning*, s. f.).

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum(Y_i - \hat{Y}_i)^2$$

Raíz del Error Cuadrático Medio (RMSE)

Error de raíz cuadrada media (RMSE) es la desviación estándar de los valores residuales (errores de predicción). Los valores residuales son una medida de la distancia de los puntos de datos de la línea de regresión; RMSE es una medida de cuál es el nivel de dispersión de estos valores residuales. En otras palabras, le indica el nivel de concentración de los datos en la línea de mejor ajuste (*Oracle® Fusion Cloud EPM Trabajo Con Planning*, s. f.-b).

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

Análisis de Residuos

Según (*RPubs - Document*, s. f.):

> "Los residuales o residuos de un modelo de regresión son la diferencia entre el valor observado de la variable dependiente y el valor esperado de la misma que se calcula mediante la ecuación de regresión."

$$\boldsymbol{e = y - \hat{y}}$$

$$\boldsymbol{\hat{y} = b_0 + b_1X_1 + b_2X_2 \ldots + b_kX_k}$$

Marco Conceptual

Inteligencia empresarial

"La inteligencia empresarial es un sistema de gestión de datos basado en datos que combina la recopilación de datos, el almacenamiento de datos y la gestión del conocimiento con el análisis para proporcionar información al proceso de toma de decisiones" (Negash, 2008, p.1).

Otros autores han afirmado lo siguiente:

> La inteligencia de negocios y el análisis se refieren al desarrollo de tecnologías, sistemas, prácticas y aplicaciones para analizar datos empresariales críticos con el fin de obtener nuevos conocimientos sobre los negocios y los mercados. Los nuevos conocimientos se pueden utilizar para mejorar los productos y servicios, lograr una mayor eficiencia operativa y fomentar las relaciones con los clientes. (Chen et. al, 2013, p.1)

Inteligencia artificial

De acuerdo con Rouhiainen (2018) menciona lo siguiente:

> En mis seminarios, intento simplificar el tema definiendo la IA como «la habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana». Pero, para brindar una definición más detallada, podríamos decir que la IA es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. Sin embargo, a diferencia de las personas, los dispositivos basados en IA no necesitan descansar y pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción

de errores es significativamente menor en las máquinas que realizan las mismas tareas que sus contrapartes humanas. (p.16)

"La IA tiene dos objetivos principales. Uno es tecnológico: usar los ordenadores para hacer cosas útiles (…). El otro es científico: usar conceptos y modelos de IA que ayuden a resolver cuestiones sobre los seres humanos y demás seres vivos" (Boden, 2016, p.5).

Seguro vehicular

"Un seguro vehicular es una herramienta que permite que los riesgos que tienen que ver con el uso de tu auto, como accidentes, robos, gastos médicos de los ocupantes y hasta gastos legales, los asuma una empresa que proporciona ese servicio" (Diners Club, 2021, p.1).

Póliza

La compañía de Seguros del Pichincha (2024) indica que:

> A rasgos generales, una póliza de seguro es el contrato entre un asegurado y una aseguradora. Este contrato traslada los riesgos a los que el asegurado está expuesto a la aseguradora a cambio del pago de una prima que garantiza la cobertura. Las pólizas de seguro establecen los derechos y obligaciones de ambas partes. El asegurado está obligado al pago de la prima del seguro. A cambio de esto, la empresa de seguros está obligada a asumir el riesgo ante los eventos establecidos en el contrato y a indemnizar al asegurado en caso de que estos se presenten. (p.1)

Prima vehicular

Según una publicación realizada por el Banco Santander (2024) sigue esta definición:

> La prima es el precio del seguro. El tomador del seguro (también llamado asegurado) está obligado al pago de la prima de acuerdo con las condiciones estipuladas en la póliza de seguro. Asimismo, la compañía aseguradora, mediante el cobro de esta, se obliga a indemnizar o satisfacer un capital, una renta u otras prestaciones convenidas en el caso de que se produzca el evento cuyo riesgo es objeto de cobertura. (p.1)

Segmentación de clientes

"El proceso por medio del cual se divide el mercado en porciones menores de acuerdo con una determinada característica, que le sea de utilidad a la empresa para cumplir con sus planes" (Bonta y Farber, 1995, p. 525).

"La manera en que una compañía decide agrupar a los clientes, con base en diferencias importantes de sus necesidades o preferencias, con el propósito de lograr una ventaja competitiva" (Hill y Jones, p. 171).

Perfil del cliente

De acuerdo con Swiderska (2024), expresa sus aportes con respecto al tema:

> Un perfil de cliente, también conocido como perfil de cliente ideal, es una descripción detallada de su cliente ideal o de un cliente anterior que ha comprado su producto o servicio. Un cliente ideal es una representación ficticia del cliente que mejor se adapta a sus productos o servicios. Un perfil de cliente contiene información demográfica como edad, sexo, ingresos y ubicación, así como información psicográfica como valores, intereses, comportamientos y puntos débiles. El objetivo de crear un perfil de cliente es conocer a fondo a su público objetivo, lo que le permitirá adaptar sus esfuerzos de ventas y marketing a sus necesidades específicas. (p.1)

> “El objetivo es identificar, describir y segmentar a los clientes en función de numerosas características y variables, basadas en sus personalidades, hábitos de compra y comportamientos” (Galic, 2024, p.1).

Centroides

Una publicación realizada por autores de Microsoft (2024), redactan lo siguiente:

> El centroide es un punto representativo de cada clúster. El algoritmo K-means asigna cada punto de datos entrante a uno de los clústeres minimizando la suma en el clúster de cuadrados. Cuando procesa los datos de entrenamiento, el algoritmo K-means comienza con un conjunto inicial de centroides elegidos al azar. (p.1)

K (número de clústeres)

“K centroides donde *k* es igual al número de clústeres elegidos para un conjunto de datos específico. Este enfoque utiliza métodos de selección aleatoria o muestreo centroide inicial” (Kavlakoglu y Winland, 2024, p.1).

Entrenamiento de datos

Los datos de entrenamiento o "training data" son los datos que usamos para entrenar un modelo. La calidad de nuestro modelo de aprendizaje automático va a ser directamente proporcional a la calidad de los datos. Por ello las labores de limpieza, depuración o "data wrangling" consumen un porcentaje importante del tiempo de los científicos de datos (De los Santos, 2023).

Test de datos

Los datos de prueba, validación o "testing data" son los datos que nos “reservamos” para comprobar si el modelo que hemos generado a partir de los datos de entrenamiento

"funciona". Es decir, si las respuestas predichas por el modelo para un caso totalmente nuevo son acertadas o no. Es importante que el conjunto de datos de prueba tenga un volumen suficiente como para generar resultados estadísticamente significativos, y a la vez, que sea representativo del conjunto de datos global (De los Santos, 2023).

Marco Legal

En el ámbito legal, este proyecto de investigación, que utiliza un modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación de clientes, requiere estar protegido y respaldado por las diferentes normativas y leyes ecuatorianas que garanticen su correcta aplicación y su no vulneración de derechos y principios. Esto debido, a que esta investigación utiliza información que puede ser vulnerable e incluso pueda afectar a terceros. Por tanto, es importante poder determinar a través de este marco legal, todas las leyes, principios, acuerdos, que puedan blindar este trabajo y que permita que este se convierta en un impulso para la innovación tecnológica tanto de la Universidad Católica Santiago de Guayaquil como para el Ecuador.

La elaboración de este marco legal se basa en la Constitución de la República del Ecuador, la cual posee los lineamientos y principios fundamentales para nuestro proyecto. En el artículo 66, numeral 19 se establece y garantiza que los ciudadanos ecuatorianos tienen el derecho a la protección de sus datos personales, incluyendo el acceso y la no manipulación y protección correspondiente de los mismos. Esto es esencial y crucial, debido a que nuestro modelo llega a manejar cierta información de clientes que puede ser un poco sensible que debe ser conservada y resguardada conforme la ley lo establece.

A su vez, el Artículo 385 de la Carta Magna ecuatoriana promueve e incita a la investigación e innovación tanto científica como tecnológica, señalando que el Estado ecuatoriano debe trabajar por fomentar el creciente y desarrollo de la ciencia, de la tecnología y los haberes ancestrales del país, ya que estos son importantes ejes estratégicos que permite la evolución y el mejoramiento de las condiciones de vida de la población. Por tanto, este artículo respalda y garantiza la viabilidad de este proyecto ante posibles críticas y opiniones, juntamente con el correcto manejo de la información previamente mencionada en el anterior párrafo.

Además, el artículo 321 de la constitución de la república ecuatoriana contempla y avala el derecho a la propiedad en sus distintas formas, manteniendo el enfoque del cumplimiento en el ámbito ambiental y social. Siento esto relevante para la validación de la protección de la propiedad intelectual de este proyecto y su correcta implementación que asegure la contribución del bienestar social y económico del Ecuador.

Por otra parte, la Ley Orgánica de Protección de Datos Personales (LOPDP) es una herramienta fundamental la realización tanto de este marco legal como de la investigación en conjunto. El articulo 1 y 10 de la misma reitera que su objetivo es asegurar el derecho a la protección personal de datos, midiendo siempre su tratamiento para garantizar los derechos de los ciudadanos y sus datos personales, permitiendo así la confidencialidad de estos, y la utilización de los datos únicamente para los fines descritos en este proyecto. Por consiguiente, se puede garantizar el cumplimiento de los artículos con respecto tratamiento de los datos personales en este proyecto, ya que este proyecto cumple con el uso correcto de variables que no afecten o puedan comprometer la integridad, reputación o generen peligro a los ciudadanos que forman parte de la base de datos a utilizar en este proyecto. Las variables

fueron meticulosamente analizadas y seleccionadas en base a el no involucramiento del ciudadano del ciudadano, si no la utilización de estas como ejemplo para la elaboración del modelo. Por tanto, variables como nombres y apellidos, ingresos, no fueron tomadas en cuenta, ya que consideramos que podrían poner en riesgo la seguridad de los prestadores de información, si no únicamente variables como sexo, provincia, edad, las cuales son comúnmente usadas para fines investigativos y que no ponen en peligro al ciudadano.

Finalmente, en cuanto al ámbito empresarial, la Política de Protección de Datos Personales de la Superintendencia de Compañías, Valores y Seguros del Ecuador, define lineamientos claros y específicos para el correcto manejo y uso de las empresas con respecto a los datos personales de sus clientes. El artículo 6 de esta política, señala que los datos recolectados deben utilizarse estricta y exclusivamente para las finalidades establecidas en este trabajo, como la segmentación y análisis de perfil de los clientes con respecto al cálculo de la prima vehicular. Cualquier muestra que viole lo señalado en este documento investigativo, deberá ser consultado con los ciudadanos y la persona que maneja la información en primera instancia. Por consiguiente, como ejecutores de este proyecto, nos hemos asegurado con el mismo este bien explicado y explicito, de tal forma que no se generen dudas que puedan llevar a una posible demanda o litigio en contra de este articulo con fines netamente investigativos, evitando así violaciones a los derechos de privacidad de los ciudadanos pertenecientes a la base de datos con la que se elabora este proyecto.

En conclusión, para el desarrollo de este trabajo investigativo se han tomado todas las precauciones necesarias que garanticen una tesis de alta calidad investigativa, y que no genere o pueda generar problemas legales en el futuro tanto a la institución como a nosotros los dos estudiantes que realizamos este trabajo. Cumpliendo así, con toda la normativa y ley

necesaria que requieren este tipo de documentos que contienen información que puede ser muy delicada y de suma confidencialidad. Permitiéndonos, cumplir a su vez con la tan mencionada ética empresarial que todo debemos de tener, promoviendo la misma y garantizando la confianza de los clientes y de las personas que depositan su información personal en nuestras manos. Demostrando así, que este proyecto no solo da todas las garantías necesarias, sino que también será un instrumento que permita formar parte del avance científico y tecnológico que requiere nuestro país.

Capítulo II

Metodología

Este trabajo de titulación es de tipo descriptivo, alimentado por componentes de carácter experimental como investigativo. Su objetivo principal es el desarrollo de un modelo que abarque los algoritmos de K Means (clasificación) y, a su vez, Regresión Lineal Múltiple (predicción), basándose en las teorías del Machine Learning. La función principal y objetiva de este modelo de aprendizaje automatizado es calcular el valor estimado de la prima vehicular y a su vez segmentar, de la forma más equitativa, los clientes dentro de los seguros vehiculares, permitiendo asi una mayor aceptación por parte del público que adquiere estos servicios.

Para la elaboración de este proyecto de investigación, se va a utilizar una base de datos de clientes asegurados perteneciente a la compañía Chevyplan S.A. Esta base posee diferentes variables como la edad, sexo, valor de prima; entendiendo asi que es una base muy completa para la ejecución del modelo, lo que permite obtener los resultados esperados de este trabajo de tesis. A partir de esto, se continua con la utilización del lenguaje de

programación R juntamente con su entorno de desarrollo integrado (IDE), Rstudio, el cual brindará las herramientas estadísticas y de visualización necesarias para le ejecución experimental de este proyecto, y donde, se va a llevar a cabo los respectivos algoritmos junto con sus validaciones que determinarán la veracidad y viabilidad del modelo, permitiendo así que pueda ser considerada futuramente como una herramienta útil para el calcular de las primas de seguro y segmentación de clientes del sector automotor (*RStudio Desktop - Posit*, 2024).

Por consiguiente, para poder realizar el respectivo análisis investigativo y experimental en la herramienta RSTUDIO, se va a preparar el respectivo entorno de trabajo y la organización de la información adquirida para la realización de este proyecto. Este proceso inicial es muy importante para la correcta ejecución de la data disponible, evitando así posibles complicaciones al momento de desarrollar el modelo. Consecuentemente, para la realización del proyecto se tomarán en cuenta cuatro etapas, que permitirán la correcta elaboración de este:

Primera etapa: Selección de variables y librería a utilizar

a) Librerías

i) library(tidyverse)

El 'tidyverse' es un conjunto de paquetes que funcionan en armonía porque comparten representaciones de datos comunes y un diseño 'API'. Este paquete está diseñado para facilitar la instalación y carga de varios paquetes 'tidyverse' en un solo paso.

ii) library(stringr)

Un conjunto de envoltorios consistente, simple y fácil de usar alrededor del fantástico paquete 'stringi'. Todos los nombres de funciones y argumentos (y posiciones) son consistentes, todas las funciones tratan con vectores de longitud cero y "NA" de la misma manera, y la salida de una función es fácil de introducir en la entrada de otra.

iii) library(tidyr)

Herramientas para ayudar a crear datos ordenados, donde cada columna es una variable, cada fila es una observación y cada celda contiene un solo valor. 'Tidyr' contiene herramientas para cambiar la forma (pivote) y la jerarquía (anidamiento y 'desanidamiento') de un conjunto de datos, convertir listas profundamente anidadas en marcos de datos rectangulares ('rectángulo') y extraer valores de columnas de cadenas. También incluye herramientas para trabajar con valores faltantes (tanto implícitos como explícitos).

iv) library(cluster)

Métodos para el análisis de conglomerados. Mucho más extenso el original de Peter Rousseeuw, Anja Struyf y Mia Hubert, basado en Kaufman y Rousseeuw (1990) "Finding Groups in Data".

v) library(factoextra)

Proporciona algunas funciones fáciles de usar para extraer y visualizar el resultado de análisis de datos multivariantes, incluidas las funciones 'PCA' (Análisis de componentes principales), 'CA' (análisis de correspondencias), 'MCA' (análisis de correspondencias múltiples), 'FAMD' (análisis factorial de datos mixtos), 'MFA' (análisis de factores múltiples) y 'HMFA' (análisis jerárquico de factores múltiples) de diferentes paquetes de R. También

contiene funciones para simplificar algunos pasos de análisis de agrupamiento y proporciona una elegante visualización de datos basada en 'ggplot2'.

vi) library (NbClust)

Proporciona 30 índices para determinar el número óptimo de clústeres en un conjunto de datos y ofrece el mejor esquema de agrupación de diferentes resultados para el usuario.

vii) library (stats)

Un documento interactivo sobre el tema del análisis estadístico básico utilizando los paquetes 'rmarkdown' y 'shiny'.

viii) library (ggplot2)

Un sistema para crear gráficos "declarativamente", basado en "La Gramática de los Gráficos". Usted proporciona los datos, le dice a 'ggplot2' cómo asignar variables a la estética, qué primitivas gráficas usar, y él se encarga de los detalles.

ix) library (lmtest)

Colección de pruebas, conjuntos de datos y ejemplos para la comprobación diagnóstica en modelos de regresión lineal. Además, se proporcionan algunas herramientas genéricas para la inferencia en modelos paramétricos.

b) Cuadro de operacionalización de las variables

Tabla 1

Cuadro de operacionalización de variables.

No.	Variable	Definición	Dimensiones	Indicadores	Escala de Medición
1	Edad	Tiempo que ha transcurrido	Cliente	Año	Escala de intervalo

		desde el nacimiento del asegurado hasta el momento.			
2	Sexo	Género del asegurado.	Cliente	Masculino/Femenino	Nominal
3	Provincia	Unidad territorial donde residente el asegurado.	Cliente	Guayas, Pichincha, Los Ríos, etc.	Nominal
4	Estado Civil	Situación legal de una persona en relación con su condición marital.	Cliente	Casado, Soltero, Divorciado, Viudo, Unión de hecho, Separado, No aplica.	Nominal
5	Tiempo Vehículo	Diferencia del año de vehículo con el año actual vigente.	Vehículo	Año	Escala de razón
6	Valor Total Vehículo	Valor por el cual se aseguró el vehículo.	Valor monetario	Expresado en dólares americanos	Escala de razón
7	Prima Mensual	Cuota a pagar al término de cada mes, durante los 12 periodos.	Valor monetario	Expresado en dólares americanos	Escala de razón

Nota. Elaboración propia. Tabla elaborada a partir de la base de datos.

Segunda etapa: Limpieza de datos

El análisis de datos atípicos y su manejo es crucial para garantizar la calidad del modelo de aprendizaje automatizado. La inclusión de datos irrelevantes o atípicos puede distorsionar los resultados, especialmente en algoritmos como el análisis clúster (e.g., K-means) y regresión lineal múltiple. A continuación, se presenta una versión mejorada del texto y se describen diferentes técnicas con sus respectivas fórmulas (Anwar, 2024).

El análisis clúster y la regresión lineal múltiple son sensibles a la presencia de datos atípicos debido a que estos distorsionan la representación real de las relaciones entre

variables. Los datos atípicos pueden surgir por errores de medición, errores humanos o condiciones extremas que no representan patrones generales. Identificarlos y manejarlos adecuadamente es esencial para obtener resultados precisos y representativos (Anwar, 2024).

a) Técnicas de limpieza de datos y detección de atípicos

i) Detección univariante

Examina una única variable para identificar valores fuera de un rango aceptable. Un criterio común es considerar como atípicos los valores que exceden tres desviaciones estándar.

Ecuación:

$$\boldsymbol{\mu \pm 3\sigma}$$

Donde:

$\boldsymbol{\mu}$: Media de los datos.

$\boldsymbol{\sigma}$: Desviación estándar.

Los valores fuera de este rango ($[\boldsymbol{\mu - 3\sigma, \mu + 3\sigma}]$) son clasificados como atípicos.

b) Imputación de valores atípicos

Si los datos atípicos se consideran errores o irrelevantes, se pueden reemplazar utilizando técnicas como:

Valor medio: $\bar{x} = \frac{1}{n}\sum_{i}^{n} 1\, x_i$ (promedio de los datos no atípicos).

Mediana: Reduce el impacto de atípicos en datos sesgados.

Interpolación: Usa relaciones conocidas entre variables para estimar el valor perdido.

c) Normalización (Min-Max Scaling)

Escala los datos a un rango específico, típicamente [0,1]. Es útil cuando los datos tienen rangos heterogéneos.

Ecuación:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Donde:

x: Valor original.

x_{min}: Valor mínimo en la variable.

x_{max}: Valor máximo en la variable.

x': Valor escalado.

d) Estandarización (Z-score Scaling)

Convierte los datos a una distribución con media 0 y desviación estándar 1. Es útil cuando los datos siguen una distribución normal.

Ecuación:

$$z = \frac{x - \mu}{\sigma}$$

Donde:

x: Valor original.

μ: Media de la variable.

σ: Desviación estándar de la variable.

z: Valor escalado.

Esto asegura que los valores estén en una escala común para algoritmos como regresión lineal o PCA.

Tercera etapa: Desarrollo del algoritmo K means

a) Algoritmo de K means

Según (Marrero, Carrizo, García-Santander, & Ulloa-Vásquez, 2021), el objetivo del análisis clúster se enfoca en agrupar objetos basándose en las características que poseen. Utilizando la función objetivo del algoritmo de k means, la cual evalúa la suma de las distancias cuadradas entre cada punto y el centroide de su respectivo clúster. Esta ecuación matemática parte de un conjunto de N observaciones de una variable aleatoria X d-dimensional $\{x_1, x_2, \ldots, x_N\}$, este algoritmo divide el conjunto de datos en un número K de clústeres conocido. Considerando a μk, con $k = 1 \ldots K$, como el conjunto de vectores d-dimensionales que representan los centros de los clústeres, los datos son asignados de manera tal que para cada uno la distancia a su respectivo centro resulta mínima en comparación con las distancias al resto de los centros (*RPubs - Final Proyect (Chap 4)*, s. f.).

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

Aquí r_{nk} es un indicador binario que refiere a cuál de los K clústeres el objeto x_n es asignado. Suponiendo su asignación al clúster k, entonces $r_{nk} = 1\ y\ r_{nj} = 0\ con\ j \neq k$. El

procedimiento es iterativo, y cada iteración implica dos pasos sucesivos correspondientes a minimizaciones sucesivas de J, en la primera fase con respecto a r_{nk} manteniendo μ_k fijo (se requiere una partición inicial), y en la segunda fase con respecto a μ_k manteniendo r_{nk} fijo. Esta optimización de dos etapas se repite hasta la convergencia, donde se obtienen los clústeres con sus respectivos clientes, los que representan los valores r_{nk} así como sus centroides, es decir, los valores μ_k, cuyo comportamiento puede ser caracterizado para la variable bajo análisis.

Finalmente, y según se expone en la literatura especializada, las reglas para seleccionar el número de grupos o clústeres son muy subjetivas porque hacen suposiciones sobre la estructura del grupo y solo funcionarían bien cuando se cumplan estos supuestos.

b) Asignación de Puntos a los Clústeres

Para cada punto x_n, calcula la distancia euclidiana $d(x_n, \mu_k)^2$ desde el punto al centroide de cada clúster μ_k (como se definió en la ecuación de distancia euclidiana). Posteriormente, se asigna el punto x_n al clúster k cuyo centroide μ_k tenga la menor distancia, es decir, asegura que cada punto se asigne al clúster más cercano.

$$r_{nk} = \{\frac{1 \; si \; k = \arg min_k || \; x_n - \mu_k||^2}{0 \; en \; caso \; contrario}$$

c) Actualización de los Centroides

Una vez que todos los puntos han sido asignados a un clúster, recalcula los centroides μ_k de cada clúster como el promedio de todos los puntos asignados a ese clúster:

$$\mu_k = \frac{1}{N_k} \sum_{n \in C_k} x_n$$

Donde N_k es el número de puntos en el clúster k y C_k es el conjunto de puntos asignados al clúster k.

d) Verificación de Convergencia

El algoritmo K-Means generalmente implica múltiples iteraciones, donde cada iteración consta de dos pasos principales: asignación y actualización. El paso de asignación implica asignar cada punto de datos al centroide más cercano, mientras que el paso de actualización recalcula los centroides en función de las asignaciones actuales. Los criterios de convergencia pueden variar, pero los umbrales comunes incluyen un número máximo de iteraciones, un movimiento mínimo de los centroides o un cambio mínimo en la función de costo general, que mide la compacidad de los clústeres (Learn Statistics Easily, 2024).

Varios factores pueden influir en la convergencia del algoritmo K-Means. La ubicación inicial de los centroides puede afectar significativamente la velocidad de convergencia del algoritmo. Una inicialización deficiente puede dar lugar a tiempos de convergencia más largos o a la convergencia a soluciones subóptimas. Se han desarrollado técnicas como K-Means++ para mejorar el proceso de inicialización, aumentando así la probabilidad de una convergencia más rápida y mejores resultados de agrupamiento (Learn Statistics Easily, 2024).

e) Distancia Euclidiana

Una vez que los clústeres han sido formados, se valida la calidad de los resultados calculando la suma de distancias entre todos los puntos y sus respectivos centroides. La distancia total se puede calcular como la suma de las distancias cuadradas entre los puntos y los centroides de sus clústeres asignados:

$$S = \sum_{n=1}^{N} ||x_n - \mu_k||^2$$

Donde μ_k es el centroide del clúster al que pertenece el punto x_n.

f) Métrica de Evaluación

Calcular la puntuación de la silueta para cada punto y para el conjunto completo de datos, es muy importante, ya que esta permite validar que tan bien se agrupan los puntos dentro de sus clústeres y qué tan bien se separan entre clústeres diferentes. Un valor cercano a 1 indica que los puntos están bien agrupados (*Silhouette_Score*, s. f.).

$$S_n = \frac{b_n - a_n}{max(a_n, b_n)}$$

Donde:

a_n es la distancia promedio de x_n a todos los puntos dentro de su propio clúster.

b_n es la distancia promedio de x_n al clúster más cercano.

El Silhouette Score total es la media de los scores de todos los puntos.

Cuarta etapa: Desarrollo de modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple, según la Universidad de Santiago de Compostela (s.f.), se expresa matemáticamente de la siguiente manera:

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \cdots + \beta_n X_n + E}$$

Donde:

y = Variable independiente

$x_1, x_2, \ldots\ x_n$ = Variables independientes

β_O = Ordenada al origen o intercepto

$\beta_1, \beta_2, \ldots\ \beta_n$= Coeficientes de regresión que representan la magnitud y dirección de la relación entre cada variable independiente y dependiente.

Para estimar coeficientes $\beta_0, \beta_1, \beta_2, \ldots\ \beta_n$ se utiliza el método de los mínimos cuadrados ordinarios, que minimza la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo (Carrasquilla-Batista et al., 2016).

$$min\beta = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$min\beta = \sum_{i=1}^{m} (y_i - (\boldsymbol{\beta_0} + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_2 X_2} + \cdots + \boldsymbol{\beta_n X_n}))^2$$

Donde:

y_i = Valor observado de la variable dependiente

$\hat{y}_i$= Valor predicho por el modelo, es decir, $\hat{y}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_2 X_2} + \cdots + \boldsymbol{\beta_n X_n}$

Este problema puede ser reformulado utilizando algebra matricial. Definimos *X* como la matriz de diseño que incluye un vector de unos para el término constante, y y como el vector de las observaciones de la variable dependiente:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{1n} \\ 1 & x_{21} & x_{22} & x_{2n} \\ 1 & x_{m1} & x_{m1} & x_{mn} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ y_m \end{pmatrix}$$

El problema de la minimización se puede reescribir en términos matriciales como:

$$m_\beta = \|y - X\beta\|^2$$

Donde:

y = Vector columna de valores observados de la variable dependiente (dimensión $m \times 1$)

X = Matriz de diseño que incluye un vector de unos para el término constante y las variables independientes $x_1, x_2, \ldots, x_n$ (dimensión $m \times (n + 1)$).

β = Vector de coeficientes a estimar (dimensión $(n + 1) \times 1$).

ε = Vector de errores

La solución óptima para β se obtiene derivando la función de costo respecto a β y resolviendo el sistema de ecuación normal:

$$\beta = (X^T X)^{-1} X^T y$$

El término de error ϵ se asume que sigue una distribución normal con media cero y varianza constante σ^2 se puede obtener como:

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

Cada coeficiente estimando $\beta_j \pm ta_{2(m-n-1)} \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}$

Donde $ta_{2(m-n-1)}$ es el valor crítico de la distribución t de student con $(m - n - 1)$ grados de libertad.

Para cada coeficiente β_j, podemos realizar una prueba de hipótesis para verificar si su valor es significativamente diferente de cero (es decir, si x_j tiene un efecto significativo sobre y). La hipótesis nula y la alternativa son:

$$H_0: \beta_j = 0, H_A: \beta_j \neq 0$$

El estadístico de prueba se calcula como:

$$t = \frac{\beta_j}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}}$$

El coeficiente de determinación R^2 mide la proporción de la variabilidad en la variable dependete que es explicada por las variables independientes:

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

Donde $\bar{y}$ es la media de los valores observados de la variable dependiente. El asjute del modelo no sólo se evalúa mediante R^2, sino también mediante la raíz del error cuadrático medio (RMSE) y la prueba F para la signifciancia global del modelo:

$$RMSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$F = \frac{\frac{SSR}{n}}{\frac{SSE}{m - n - 1}}$$

Donde SSR es la suma de los cuadrados de regresión y SSE es la suma de cuadrados de los errores. Una vez estimados los coeficientes β, el modelo lineal ajustado sería:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Donde $\hat{y}$ representa la predicción de la variable dependiente basada en las variables independientes y sus coeficientes estimados. Por lo que es importante verificar la calidad del modelo utilizando métodos de diagnóstico como:

a) **Gráficos de residuos**: Para identificar cualquier patrón no explicado.
b) **Prueba de significancia de coeficientes:** Utilizando el estadístico t para determinar si los coeficientes son significativamente diferentes de cero.
c) **Validación cruzada:** Para asegurar que el modelo generaliza bien a datos no observados.

Capítulo III

Análisis de resultados

En base a la metodología expuesta previamente, primero se realizar el script en R Studio. Se importa la base de datos denominada “Seguros”:

```
Seguros <- read.csv2("../Desktop/Seguros.csv")
```

La función *read.csv2()*: se utiliza para leer archivos cuyos valores se encuentran separados por comas.

Para la elaboración del modelo tanto de regresión como de k-means para el cálculo de la prima vehicular y la segmentación de clientes, se instalan las siguientes librerías:

```
library(class)

install.packages("class")
```

```
library(caret)

install.packages("caret")

library(tidyverse)

install.packages("tidyverse")

library(cluster)

install.packages("factoextra")

library(factoextra)

install.packages("NbClust")

library(NbClust)
```

De acuerdo con la teoría de K-means, se necesita que todas las variables a utilizar sean de naturaleza numérica.

Por lo tanto, se emplea la función *factor()*: la cual factoriza las variables para representar datos categóricos con un número finito de valores, como se demuestra a continuación con la variable sexo:

```
Seguros$SEXO <- factor(Seguros$SEXO)
```

Una vez hecho esto, se aplica la función *as.numeric ()*: útil para transformar los factores en números.

```
Seguros$SEXO <- as.numeric(Seguros$SEXO)
```

Se replica este procedimiento con las variables restantes que son provincia y estado civil.

```
Seguros$PROVINCIA <- factor(Seguros$PROVINCIA)

Seguros$PROVINCIA <- as.numeric(Seguros$PROVINCIA)

Seguros$ECIVIL <- factor(Seguros$ECIVIL)

Seguros$ECIVIL <- as.numeric(Seguros$ECIVIL)
```

De allí, con la función na.omi (): se omiten los valores faltantes que contiene la base para trabajar con un conjunto de datos limpios y realizar un buen análisis.

```
Seguros <- na.omit(Seguros)
```

Ahora bien, se empieza con la clasificación de los datos con el uso de la función set.seed(123), cuyo objetivo final es reproducir una secuencia de números aleatorios.

Luego, los datos se dividen en 80% Entrenamiento y 20% Prueba, los cuales se centran alrededor de la prima mensual, variable que se busca predecir, como se demuestra a continuación:

```
trainIndex <- createDataPartition(Seguros$Prima_Mensual,

p=.8, list = FALSE)

train_data <- Seguros[trainIndex,]

test_data <- Seguros[-trainIndex]
```

trainIndex contiene los datos originales para el entrenamiento, la función *createDataPartition* proviene de la paquetería “caret” utilizada para dividir el conjunto de datos, *Seguros$Prima_Mensual* es la variable objetivo, *p=0.8* indica que el 80% de los datos estarán destinados para entrenamiento y el 20% para prueba, *list = FALSE* indica que el

resultado debe reflejarse como un vector y no una lista. Y *train_data* serán los datos de entrenamiento mientras que *test_data* serán los de prueba.

Después se procede con el escalamiento de datos, tanto para los datos de Entrenamiento (*SKtraindata*) como para el de Prueba (*SKtestdata*), la función *scale ():* sirve para escalar y centrar las columnas de la matriz, consiguiendo que su media sea 0 y su desviación estándar de 1, se asegurará que las variables sean equitativas y no afecten al modelo:

```
SKtraindata <- scale(train_data)

SKtestdata <- scale(test_data)
```

A continuación, se trabajará con los datos de Entrenamiento para determinar el número óptimo de clústeres; para ello, se recurre a la función *fviz_nbclust ():* perteneciente a la paquetería "factoextra", donde se generan gráficos para identificar la cantidad adecuada de agrupamiento mediante distintos métodos.

El primero a usar es el método silueta:

```
fviz_nbclust(SKtraindata, kmeans, method = "silhouette")
```

El cual mide la distancia de los puntos dentro de un mismo clúster y la distancia con los puntos de clústeres vecinos. El coeficiente de cada punto se encuentra dentro de un rango desde -1 a 1, donde 1 significa que los clústeres están bien agrupados y -1 que los puntos están más adyacentes a los clústeres cercanos.

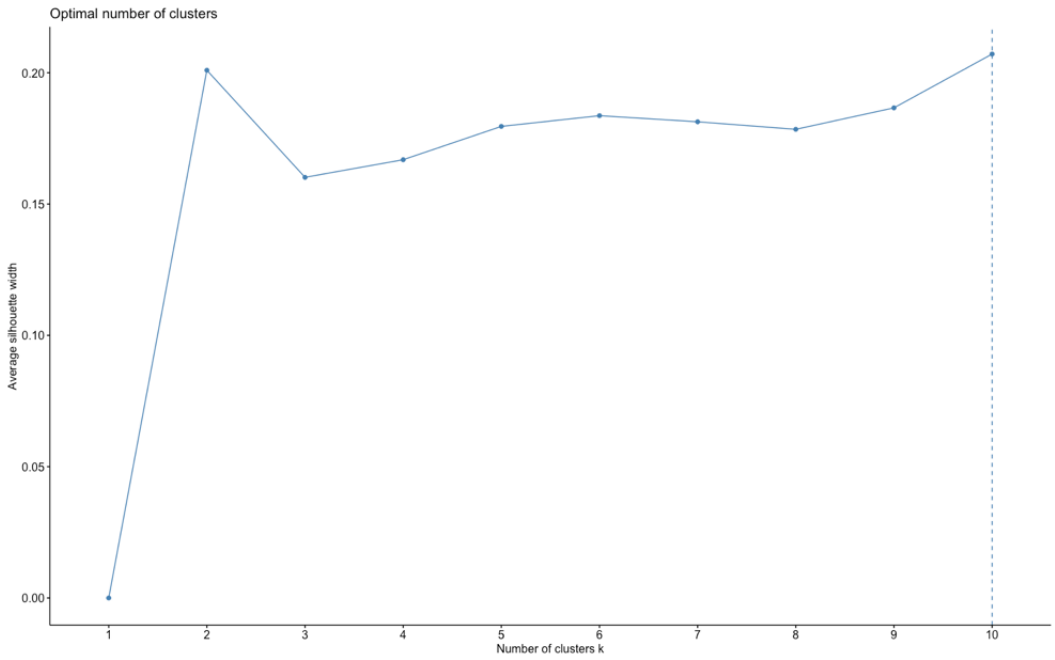


Gráfico 1. Método de silueta. Elaboración propia.

Se puede observar que el pico más alto dentro del gráfico indica que el número óptimo de clúster es 2.

Ahora sigue el método de codos:

```
fviz_nbclust(SKtraindata, kmeans, method = "wss")
```

El mismo consiste en el cálculo de la suma de cuadrados dentro de cada clúster; de esta manera, se puede observar la proximidad de cada punto a su centroide. A medida que aumentan los números de clústeres, los datos se compactan más y reduce la variabilidad dentro de los mismos.

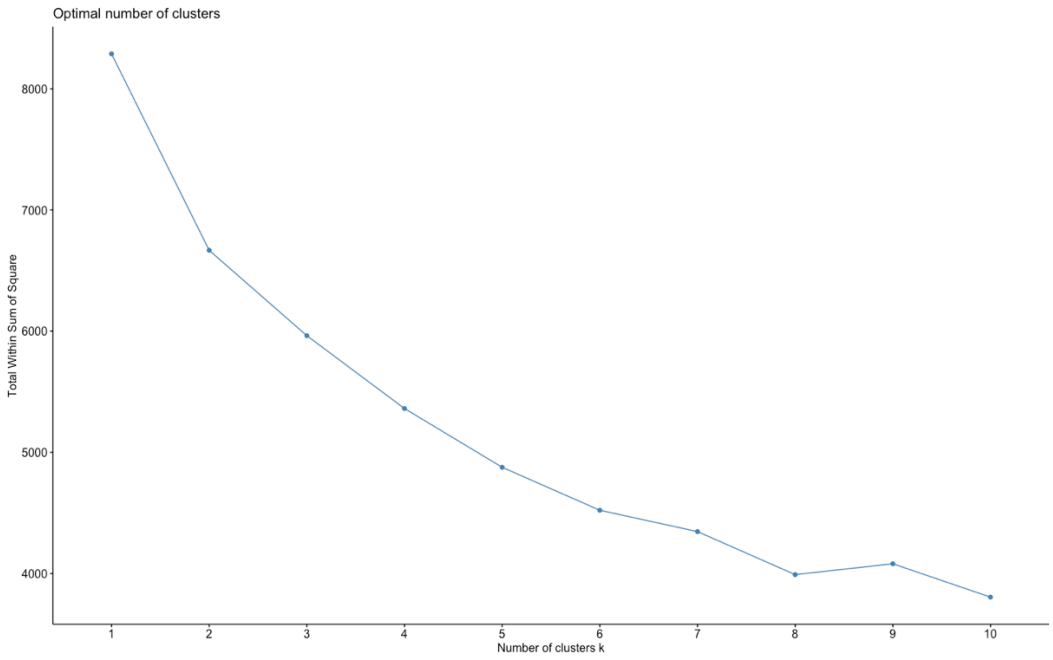


Gráfico 2. Método de codos. Elaboración propia.

Se puede identificar que el punto de inflexión es en el 2, lo cual indica que ese el número óptimo de clúster para el modelo.

Y por último, se utiliza el método de brecha:

```
fviz_nbclust(SKtraindata, kmeans, method = "gap_stat")
```

Se basa en la comparación entre la dispersión dentro de los clústeres reales y la dispersión esperada de un modelo nulo generado de forma aleatoria.

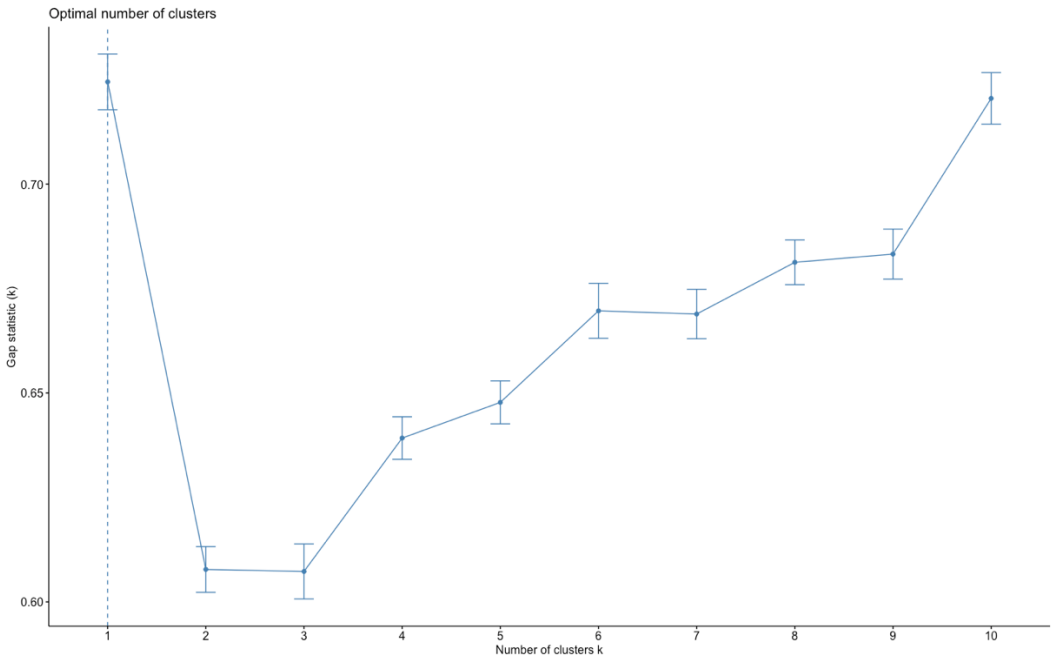


Gráfico 3. Método de brecha. Elaboración propia.

Según el gráfico, el número óptimo de clústeres es 1 pero la varianza se corta en 2, es decir, puede ser tanto 1 como 2 clústeres.

En suma, el método de silueta y el método de codos indican que el número óptimo de clústeres es 2; mientras que, el método de brecha, indica que puede ser 1 o 2. Siguiendo la regla de la mayoría, se determinó trabajar con 2 clústeres. Con esto en mente, se procede con el cálculo del K-medias con 2 centros.

```
K2 <- kmeans(SKtraindata, centers = 2, nstart = 25)
```

El desglose de la función anterior es la siguiente: *K2* el resultado del modelo, *kmeans()* ejecuta el algortimo K-means, *SKtraindata* es el conjunto de datos escalados, "*centers=2*" es el número de clústeres y "*nstart=25*" significa que el algoritmo se ejecutará 25 veces y seleccionará el mejor agrupamiento.

La salida del modelo da como resultado:

```
K-means clustering with 2 clusters of sizes 408, 777

Cluster means:
         EDAD        SEXO   PROVINCIA       ECIVIL Tiempo_Vehiculo Valor_Total_Vehiculo Prima_Mensual
1  0.02833764  0.11242774 -0.04443274 -0.018728006      -0.6704039             1.049586      1.019053
2 -0.01488000 -0.05903542  0.02333148  0.009834011       0.3520268            -0.551134     -0.535101
```

Gráfico 4. K2. Elaboración propia.

Primero, se aprecia que el clúster 1 contiene 408 observaciones, mientras que el clúster 2 contiene 777 observaciones. Después, se obtiene el promedio por variable dentro del espacio escalado, donde en términos estandarizados se interpreta que:

a) El promedio para la variable *Edad* en el clúster 1 es de 0.0283 y para el clúster 2 es de -0.0148, por el hecho de ambos estar cerca del 0, significa que no existe mayor diferencia.
b) La variable *Tiempo_Vehículo*, muestra un promedio de -0.670 (inferior al promedio general) en el clúster 1, mientras que en el clúster 2, demuestra un promedio de 0.352, siendo superior al promedio general.
c) Para la variable *Valor_Total_Vehículo*, el promedio del clúster 1 es de 1.049, sugiriendo que los vehículos tienden a ser más costosos y para el clúster 2 es de -0.551, sugiriendo que son más baratos.
d) En cuanto a la variable *Prima_Mensual*, el promedio del clúster 1 es de 1.019, agrupando a clientes con primas más altas del promedio general mientras que el del clúster 2 es de -0.535, sugiriendo lo contrario.

Evidentemente, las dos últimas variables son un factor diferenciador significativo entre los clústeres.

Avanzando con el script, se generan los gráficos del K-medias, empezando por el plot de entrenamiento:

```
PTrainOC <- fviz_cluster(K2, data = SKtraindata)

PTrainOC
```

PTrainOC es el nombre del gráfico, la función *fviz_cluster()* es para graficarlo, *K2* es el modelo calculado previamente, *data = SKtraindata* es el conjunto de datos escalados.

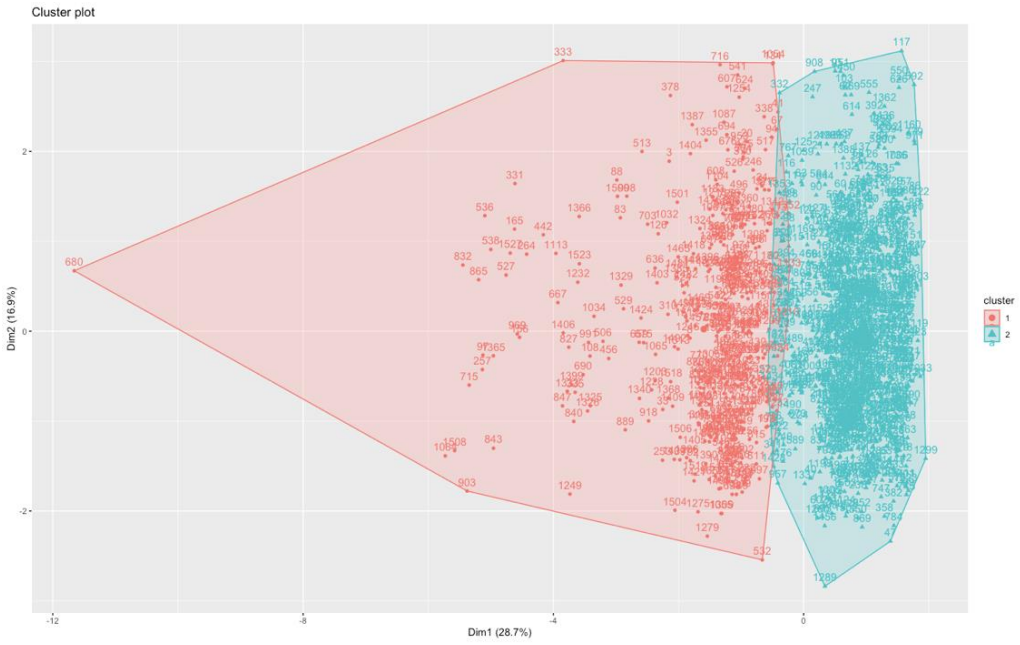


Gráfico 5. Plot Train Original Cluster. Elaboración propia.

Para empezar, se puede denotar que existen dos claros segmentos, y bajo el lineamiento de que la variable de salida es la prima mensual, se puede interpretar que el grupo 1 (rojo) es más correlacionado que el grupo 2, dando a entender que posiblemente albergue primas más bajas. Por otro lado, el clúster 2 (azul) corresponden a primas más caras. Adicionalmente, se puede deducir que la dimensión 1 contenga variables económicas como el valor del vehículo, lo cual influye en el precio de la prima.

El gráfico migratorio se obtiene de la siguiente forma:

```
PTrainMC <- fviz_cluster(K2, data = SKtraindata,

ellipse.type = "euclid",

repel = TRUE, star.plot= TRUE)

PTrainMC
```

PTrainMC es el nombre del gráfico, la función *fviz_cluster()* es para graficarlo, *K2* es el modelo calculado previamente, *data = SKtraindata* es el conjunto de datos escalados, *ellipse.type="euclid"* grafica un círculo que representa la distancia euclidiana desde el centro, *repel = TRUE* ajusta las etiquetas dentro del gráfico y *star.plot= TRUE* grafica las líneas que entrelazan cada punto con el centroide de su clúster.

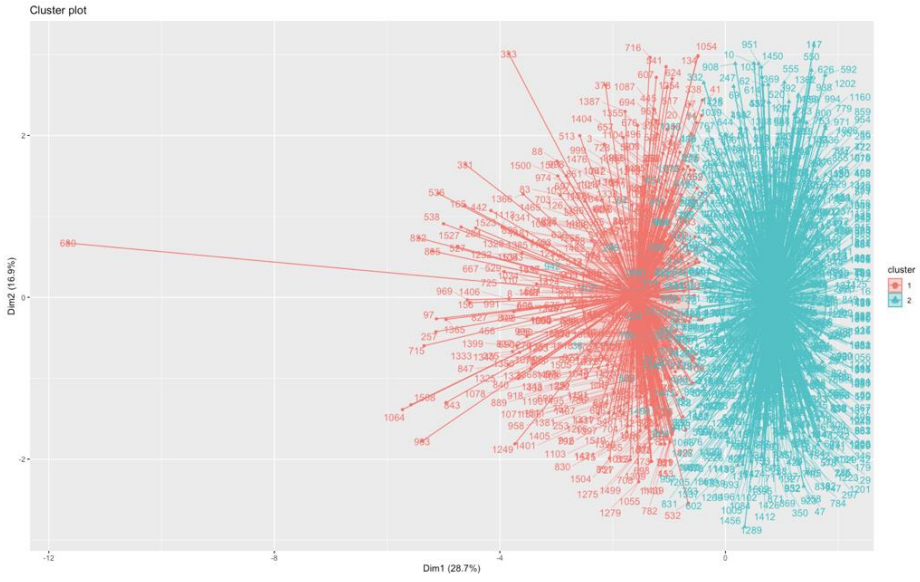


Gráfico 6. Plot Train Migratory Cluster. Elaboración propia.

Se puede interpretar que el grupo 2 (azul) busca integrarse al grupo 1 (rojo), esto se evidencia en las líneas alejadas del centroide del grupo 2. En síntesis, los del grupo más caro quieren obtener primas más baratas.

Y el gráfico del intercepto se alcanza de la siguiente manera:

```
PTrainIC <- fviz_cluster(K2, data = SKtraindata,

ellipse.type = "norm")

PTrainIC
```

PTrainIC es el nombre del gráfico, la función *fviz_cluster()* es para graficarlo, *K2* es el modelo calculado previamente, *data = SKtraindata* es el conjunto de datos escalados, *ellipse.type="norm"* grafica los elipses en base a una distribución normal multivariada.

Gráfico 7. Plot Train Intercept Cluster. Elaboración propia.

Nuevamente, el gráfico da a entender que el grupo 2 (azul) quiere ser parte del 1 (rojo). Además, se puede notar una superposición entre los eclipses, demostrando que existen características de los clientes de ambos grupos que se comparten entre sí.

Acto seguido, se procede con el mismo procedimiento de clusterización pero con los datos de prueba:

```
Kt2 <- kmeans(SKtestdata, centers = 2, nstart = 25)
```

Como resultado se obtuvo que el clúster 1 contiene 7701 observaciones, mientras que el clúster 2 contiene 1481. En cuanto a la variable objetivo, prima mensual, el clúster 1 tiene un promedio de -0.414 y el clúster 2 tiene un promedio de 2.157.

```
K-means clustering with 2 clusters of sizes 7701, 1481

Cluster means:
        [,1]
1 -0.4149708
2  2.1577922
```

Figura 19. *Cluster means. Elaboración propia.*

En general, se observa que existe coherencia entre los resultados del conjunto de datos de entrenamiento y el conjunto de datos de prueba, como se evidencia en la tabla adjunta:

Tabla 2

Coherencia entre conjunto de datos

	Datos de entrenamiento	**Datos de prueba**
Clúster 1	Agrupa la minoría de observaciones (408) y contiene primas altas (1.019053)	Agrupa la mayoría de observaciones (7701) y contiene primas bajas (-0.4149708).
Clúster 2	Agrupa la mayoría de observaciones (777) y contiene primas bajas (-0.535101).	Agrupa la minoría de observaciones (1481) y contiene primas altas (2.1577922).

Nota. Elaboración propia. Tabla elaborada a partir de los resultados de "Cluster Means" en R-Studio.

El patrón de que el clúster con la mayoría de las observaciones contiene las primas más bajas se mantienen en ambos conjuntos de datos. De igual forma ocurre con las primas, existe consistencia entre los datos, a pesar de que las primas son mucho más altas en el conjunto de prueba. Se confirma que el modelo generaliza adecuadamente.

Posteriormente, se pasan los datos a la base:

```
Seguros$class <- NA

Seguros$class[trainIndex] <- K2$cluster

Seguros$class[-trainIndex] <- Kt2$cluster
```

Para iniciar, se debe crear la columna “class” dentro de la base original, la cual contiene valores vacíos. Luego, dichos valores son llenado mediante la asignación de las clasificaciones provenientes de los datos de Entrenamiento y Prueba. Es así, como concluye la primera parte correspondiente a K-Means.

Tomando en cuenta lo establecido anteriormente en la metodología de este trabajo investigativo, se continuará con el algoritmo RLM, el cual será útil en el desarrollo del Modelo de Aprendizaje Automatizado que calculará la prima de seguro de un vehículo. Como establece la teoría, el modelo de RLM depende de una variable ‘y’ o también conocida como variable de salida, la cual será el punto de partida de este algoritmo, ya que esta es ejemplo al cual debemos predecir y aproximar nuestra ‘variable resultada’ del algoritmo, obteniendo así el error cuadrático más bajo posible que permita determinar que lo predicho en esta variable sea correcto. Por lo tanto, se planta la semilla:

```
set.seed(2018)
```

Esta función de R permitirá que continuamente haya datos aleatorios al momento de realizar la partición de datos, asegurando así que cualquier operación dentro del script sea reproducible. Cabe destacar, que no existe ninguna diferencia en la semilla del algoritmo de K-Means y esta, netamente al cambiar de semilla, la ejecución aleatoria de los datos comienza desde otro punto de la base. Esto garantiza que el modelo sea robusto, que el modelo no es producto de una configuración inicial y que tendrá el mismo comportamiento general en los distintos escenarios que se le presenten, reduciendo así el riesgo de sobreajuste del modelo, o que simplemente no se pueda obtener los resultados esperados.

Consecuentemente, se procede con la partición de datos para el entrenamiento y prueba del modelo:

```
Entrenamiento <- createDataPartition (Seguros$Prima_Mensual,
p=.8, list = FALSE)
```

Utilizamos la función createDataPartition, donde al igual que en el algoritmo de K medias, se hará la partición de datos. El p=8 representa que se asignará el 80% de los datos de la base de seguros para el conjunto de entrenamiento. Permitiendo así, que, con el entrenamiento, el modelo pueda aprender patrones y relaciones subyacentes entre las variables dependientes con la variable independiente que en este caso es 'Seguros$Prima_Mensual', respaldando así la precisión en la predicción de la variable resultante. Finalmente, se utiliza la función list = FALSE, para crear un vector que devuelva los índices de las filas directamente, sin convertirlos en lista. De tal forma, una vez explicado ejecutado cada una de las funciones de R que interfieren en el proceso de partición de datos, se puede continuar con el siguiente paso en el desarrollo del modelo, que se encuentra en la siguiente línea de código, y es la ejecución de ese 80% de datos para el entrenamiento del modelo:

```
Reg <- lm(Prima_Mensual~.,data = Seguros[Entrenamiento,])
summary(Reg)
```

En este código se puede observar la utilización de funciones como lm, la cual crea un modelo de regresión lineal, ajustándolo a los datos proporcionados, los cuales son (Prima_Mensual~.,data = Seguros[Entrenamiento,]). Finalmente, el modelo se guarda en el objeto'Reg', que contiene los coeficientes estimados, los residuales y las estadísticas

generales del modelo entrenado. Se visualiza lo mencionado anteriormente a través de la función summary(Reg), la cual muestra lo siguiente:

```
Residuals:
    Min      1Q  Median      3Q     Max
-74.118  -5.508  -1.251   3.960  83.367

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.776e+01  2.651e+00   6.699 3.25e-11 ***
EDAD                -3.175e-03  2.374e-02  -0.134   0.8936
SEXO                -8.638e-01  6.029e-01  -1.433   0.1522
PROVINCIA           -1.340e-01  5.822e-02  -2.301   0.0215 *
ECIVIL              -1.254e-01  1.508e-01  -0.831   0.4060
Tiempo_Vehiculo     -1.596e+00  5.176e-01  -3.083   0.0021 **
Valor_Total_Vehiculo 3.065e-03  4.541e-05  67.493  < 2e-16 ***
class               -1.223e+00  7.601e-01  -1.609   0.1080
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.05 on 1177 degrees of freedom
Multiple R-squared:  0.8596,    Adjusted R-squared:  0.8587
F-statistic:  1029 on 7 and 1177 DF,  p-value: < 2.2e-16
```

Gráfico 8. Valores residuales del modelo de regresión múltiple. Elaboración propia.

Resumiendo, se observan las variables que consideradas importantes para la predicción, las cuales están marcadas por uno o más asteriscos, son únicamente las de PROVINCIA, Tiempo_Vehiculo y Valor_Total_Vehiculo, estas variables serían las más que significativas y que tendrían mayor impacto o repercusión en la Prima_Mensual. Además, a través de los residuales máx. y min., el modelo considera que puede haber algunos o pocos datos aberrantes o extremos que pueden estar afectando a la precisión del modelo, a pesar de que el mismo es sólido, ya que posee un error cuadrático de 0,8 que permite entender que está captando casi completamente la variabilidad de la variable ‘y’ en el modelo y por ende, el mismo está teniendo una correcta predicción. Sin embargo, se requieren mayores

visualizaciones para observar que el modelo de RLM este correctamente ajustado, por ello, se aplica lo siguiente:

```
boxplot(Reg$residuals)
```

A través de esta función, se puede elaborar un boxplot o también conocido como 'Diagrama de Caja', que permitirá visualizar las distancia entre los datos y la media. De tal, forma que, si la media está bien ubicada entre en el centro, quiere decir que los datos están bien distribuidos y equilibrados dentro del modelo, permitiendo así tener confianza en el modelo y en los datos, ya que no existen sesgos en las variables. Por tanto, al ejecutar la función, resulta lo siguiente:

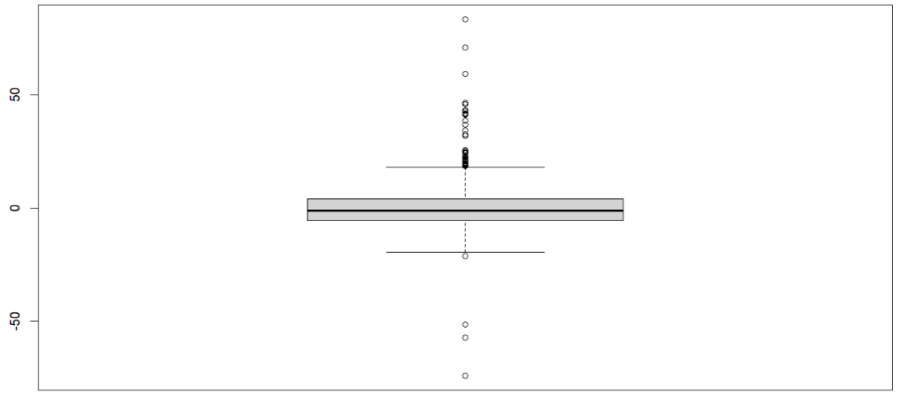


Gráfico 9. Boxplot de residuales. Elaboración propia.

Si bien es cierto, se puede apreciar que la media residual esta casi que perfectamente ubicada en el 0, también se observan diferentes datos extremos que pueden estar afectando a la confianza e incluso a la precisión en cuanto a la predicción y evolución de residuales en el modelo. Por consiguiente, para poder evaluar mayormente si los valores si van a predecir correctamente con estos datos de entrenamiento, se procede con la siguiente función:

```
Prediccion <- predict(Reg,Seguros[-Entrenamiento,])
```

A través de esta función, guardado en el objeto 'Prediccion', se evalúan los valores predichos para la variable dependiente en función a las variables independientes y el respectivo modelo ajustado. Por tanto, una vez ejecutado este código, se da paso a la siguiente línea de código que permitirá visualizar diferentes gráficos de diagnósticos del modelo de RLM:

```
par(mfrow=c(2,2))

plot (Reg)
```

A través de la función mencionada anteriormente, donde 'par' concede la modificación de los parámetros de los gráficos a mostrar y 'mfrow' como se dividirá la ventana gráfica, que en este caso se dividide en (2,2), dos filas y dos columnas, que permite exponer cuatro gráficos a través de la función 'plot', los cuales muestran lo siguiente:

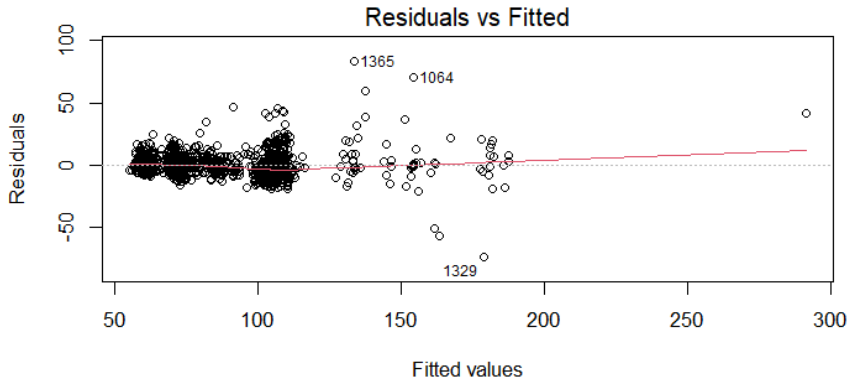


Gráfico 10. Residuos vs Ajuste. Elaboración propia.

A continuación, en el primer gráfico de valores residuales contra valores ajustados, la función de este gráfico es verificar que los residuos se están distribuyendo de manera

aleatoria y poseen una variabilidad o varianza constante. Por consiguiente, lo ideal sería que los puntos caigan de manera aleatoria por ambos lados de la línea de cero, en estos puntos no se deben generar patrones reconocibles. Por lo cual, tomando en cuenta la teoría, en el primer gráfico se pueden ver que los puntos se están distribuyendo de manera aleatoria, y además no forman ningún patrón, por ende no muestra linealidad; sin embargo, sí se puede observar que existen algunos datos alejados de la línea del cero, por ende el gráfico muestra que existen outliers, como el punto '1064' que deben ajustarse, para que así no se afecte la homocedasticidad del modelo, y a su vez no exista sesgo y problemas de predicción del mismo. Por consecuente, se debe ajustar el modelo para poder eliminar estos valores extremos que afectan tanto al modelo como a su capacidad de generalización.

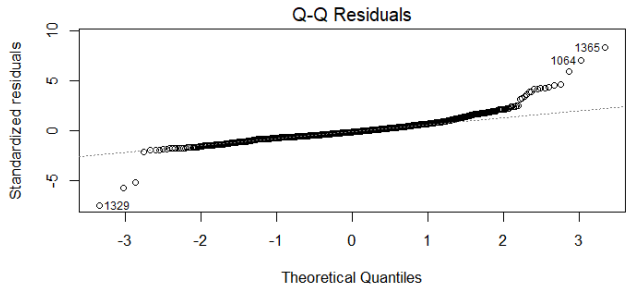


Gráfico 11. Cuantil-Cuantil. Elaboración propia.

En segundo gráfico se trata del Q-Q Residuales o también llamado Cuantil-Cuantil residuales, el cual, según la teoría consiste en que se pueda observar visiblemente que los datos siguen una distribución específica y cumplen una simetría (distribución normal). Si bien es cierto que los datos están visiblemente alineados y simétricos, lo cual establece que siguen un patrón determinado y a su vez poseen linealidad, se puede verificar que en las colas existen datos que se encuentran dispersos, por ende, el modelo sugiere que existen valores

atípicos que no están siguiendo o teniendo una distribución normal, y que estos pueden estar afectando directamente a los coeficientes o valores residuales del modelo, generación a su vez una afectación en la robustez del mismo.

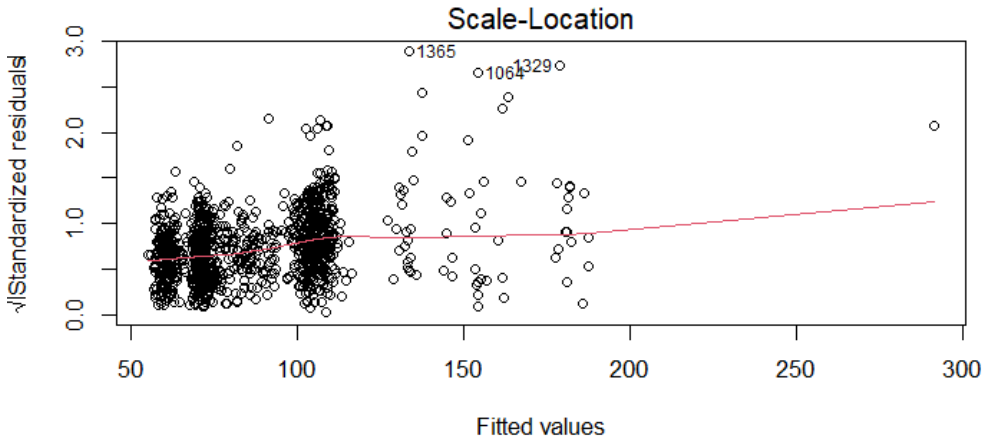


Gráfico 12. Scale Location. Elaboración propia.

En el tercer gráfico, llamado Scale-Location, se muestra si los residuos se distribuyen de forma uniforme a lo largo de los rangos de los predictores. De esta forma, se puede comprobar el supuesto de varianza igual (homocedasticidad). Es conveniente que vea una línea horizontal con puntos distribuidos de forma uniforme y aleatoriamente. (UVA Library, s. f.).

Ahora bien, se observa que la gran mayoría de los puntos están dispersos aleatoriamente y cerca de la línea cero, lo cual facilita entender que la varianza de los puntos es constante; no obstante, se evidencia una ligera curvatura en la línea roja, que puede establecer la existencia de posible presencia de heterocedasticidad en los valores ajustados

altos del gráfico. Entendiendo de tal manera, que también existe outliers que pueden estar afectando a la linealidad y al ajuste del modelo.

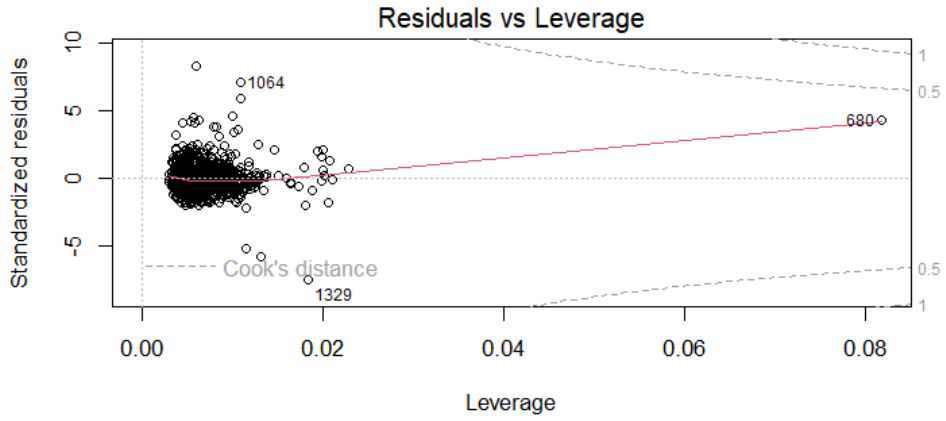


Gráfico 13. Residuos Vs Apalancamiento. Elaboración propia.

En el último y cuarto gráfico, llamado Residuos vs. Apalancamiento, se detallan qué casos o puntos pueden estar afectando al modelo debido a que pueden estar fuera o superando límites predictores del modelo. Por eso, entendiendo la teoría, se observa la existencia de valores atípicos que están distantes del centro, donde se encuentran los predictores, por tanto, puede afectar significativamente en los coeficientes residuales del modelo, a pesar de que no hayan superado los límites del gráfico.

Finalmente, en cuanto a los gráficos, en su mayoría muestran que los datos están distribuidos aleatoriamente, y a su vez la relación entre las variables es sobre todo lineal. Sin embargo, existen valores atípicos que pueden estar afectando tanto los coeficientes residuales mostrados en el entrenamiento del modelo. Por ende, al tener estos hallazgos, se procede con la optimización del modelo utilizando la siguiente función:

```
Prediccion2 <- predict(RegOp,Seguros[-Entrenamiento,
```

-c(1,2,4)])

En esta línea de código se incluye únicamente las variables de mayor significancia, que se mostraron en los resultados de la anterior predicción, obteniendo los siguiente:

```
Residuals:
    Min      1Q  Median      3Q     Max
-74.648  -5.605  -1.267   4.073  82.826

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.564e+01  2.182e+00   7.164 1.37e-12 ***
PROVINCIA           -1.336e-01  5.778e-02  -2.311  0.02098 *
Tiempo_Vehiculo     -1.608e+00  5.159e-01  -3.118  0.00187 **
Valor_Total_Vehiculo 3.067e-03  4.539e-05  67.570  < 2e-16 ***
class               -1.121e+00  7.564e-01  -1.482  0.13873
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.05 on 1180 degrees of freedom
Multiple R-squared:  0.8592,    Adjusted R-squared:  0.8588
F-statistic:  1801 on 4 and 1180 DF,  p-value: < 2.2e-16
```

Gráfico 14. Valores residuales ajustados del modelo de regresión múltiple. Elaboración propia.

Al momento de comparar estos resultados con los resultados del cuadro anterior del modelo de regresión, se puede concluir que, tras la eliminación de variables no significativas o redundantes del modelo, los residuales no cambian de manera significativa en su distribución ni en su error estándar promedio, lo que indica que las variables eliminadas no aportan información relevante al modelo. Además, los valores de R2 y R2 ajustado permanecieron casi constantes, lo que confirma que el modelo reducido conserva una capacidad explicativa sólida y estadísticamente significativa. Esta reducción mejora la parsimonia del modelo sin comprometer su calidad predictiva. Una vez obtenido estos resultados, se volverá a realizar los gráficos que nos permitan entender, si se están

cumpliendo todos los parámetros que me permiten validar que el modelo es robusto, es lineal, y cumple con homocedasticidad. Primeramente, se vuelve a emplear el gráfico de boxplot para poder visualizar los residuales:

```
par(mfrow=c(1,1))

boxplot(RegOp$residuals)
```

La primera función ayuda a deshacer la partición de la pantalla y se muestra únicamente un gráfico. Una vez ejecutado esta función, se continua con el boxplot, el cual muestra la siguiente gráfica:

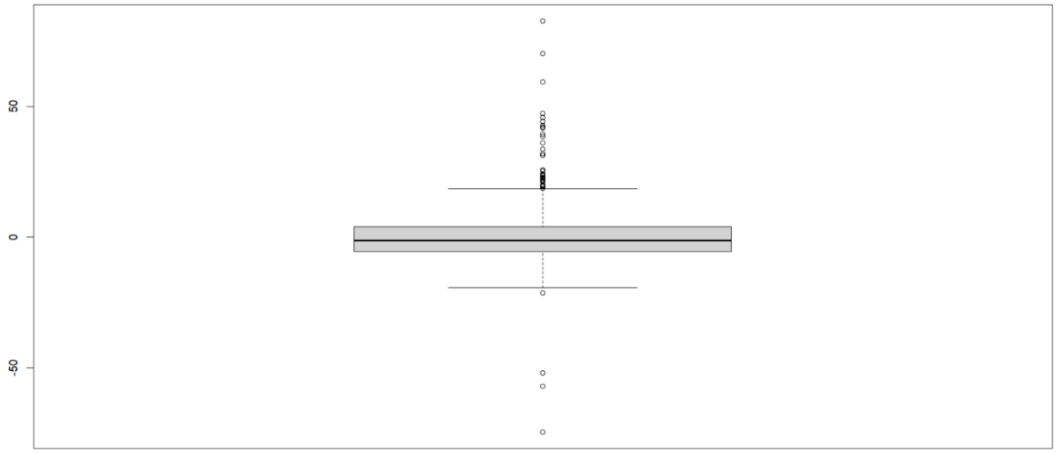


Gráfico 15. Boxplot con variables ajustadas. Elaboración propia.

En esta visualización, se mantiene la mediana en el centro, entonces, hay simetría en los datos; sin embargo, a pesar de que han disminuido los errores atípicos, permitiendo así que estos influyan menos en la precisión del modelo, sigue existiendo algunos puntos que posiblemente estén prediciendo con dificultad o con valores errores, pero que posiblemente no afecte la generalización del modelo. Por ende, al haber entrenado los datos con las

variables de mayor significancia, se procede con la predicción de la tendencia con los datos de prueba a través de la siguiente línea de código:

```
Prediccion2 <- predict(RegOp,Seguros[-Entrenamiento,

-c(1,2,4)])
```

Se aplica la misma función de la predicción anterior, únicamente con la diferencia que se excluyen las variables que no son significantes como la variable 'Sexo' y consecuentemente se utiliza 'Entrenamiento', para poder ejecutar la predicción con los datos de prueba. Una vez realizada la predicción, se vuelve a utilizar nuestras gráficas que posibilitará medir los residuales contra los datos predichos, por lo cual se utiliza la misma función que se utilizó anteriormente para poder graficar las cuatro visualizaciones, y que manifiesten las siguientes gráficas:

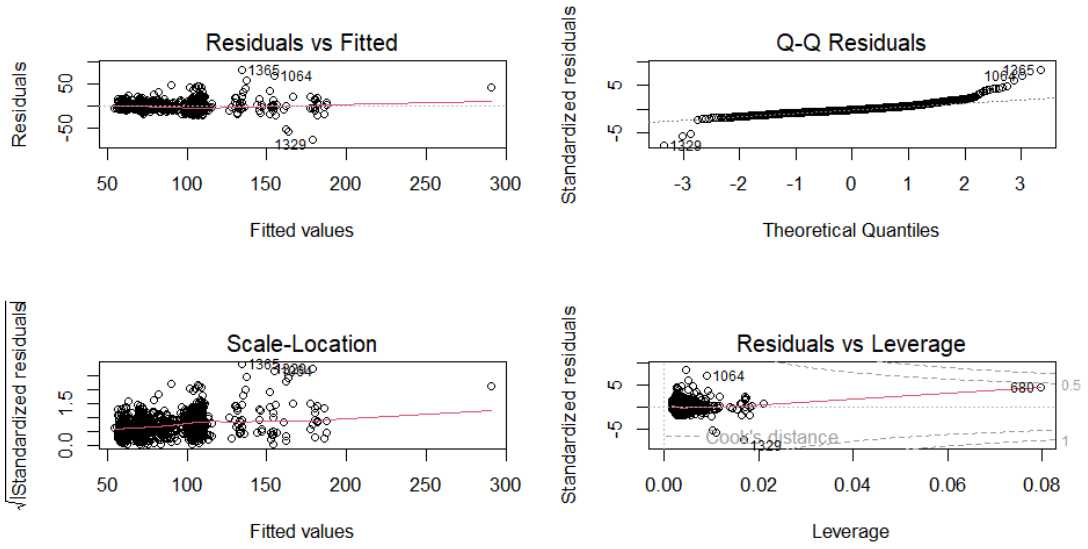


Gráfico 16. Valores residuales. Elaboración propia.

Al analizar y comparar las gráficas actuales de residuos con las gráficas anteriores, se puede determinar que el ajuste realizado, a través de la reducción de variables, no ha inferido mayormente en el resultado. En el primer gráfico, los valores residuales siguen estando distribuidos, en su gran mayoría, de manera aleatoria alrededor de la línea 0, y se mantiene los muy pocos los valores extremos, destacando observaciones como 1365, 1064 y 1329.

En el segundo gráfico, se muestra un comportamiento muy similar con el modelo anterior, teniendo un muy acertado ajuste en el rango medio formando una linealidad, que garantizaría en su mayoría la normalidad que hay en los datos. Sin embargo, aún estas presentes los valores atípicos en las colas, que en primera instancia si han disminuido, pero aún se encuentran presentes.

En el tercer gráfico, al igual que en el gráfico anterior de Scale-Location, se muestra la homocedasticidad en los datos, manteniendo una varianza similar, a pesar de una pequeña o ligera tendencia en ciertos datos atípicos que generan una variabilidad en la varianza debido a su distancia de la recta. Finalmente, en el cuarto gráfico, la gran mayoría de datos se encuentran dentro de los intervalos predicción, a excepción de los datos 680 y 1329, que continúan distantes.

En resumen, la reducción de variables permitió una pequeña mejoría en el comportamiento residual, aportando a la mejora en la presión del modelo. Sugiriendo que las variables que poseían poca significancia no aportaban la información esencial del modelo para el ajuste, por ende, permitió mejoras pequeñas en los residuos, que certifican aún más la solidez del modelo a comparación del anterior.

Por consiguiente, una vez ya ajustado el modelo, tratando de reducir los residuos hasta el mínimo, se procede con la predicción de la variable resultando con respecto a las variables independientes, y tomando de base la variable dependiente, para esto se utilizó las siguientes líneas de código:

```
Seguros[-Entrenamiento,"Prediccion"] <- predict(RegOp,

Seguros [-Entrenamiento,

-c (1,2,4)])

Seguros [Entrenamiento,"Prediccion"] <- predict(RegOp,

Seguros [Entrenamiento,

-c (1,2,4)])
```

A través de las funciones señaladas se establece la predicción generada en una columna que se llama ‘Prediccion’, tanto para los datos de entrenamiento como de prueba, dentro de la base de Seguros. Para ambos conjuntos de datos, se utilizó únicamente las variables ‘Poliza_vehicular’, ‘Provincia’ y ‘Valor_del_vehiculo’, excluyendo el resto de las variables, ya que como se mencionó anteriormente, las variables como ‘Sexo’, ‘Edad’, ‘Estado civil’ y ‘class’, no daban mayor aporte para la predicción del modelo, y a su vez, sin estas, existía una ligera mejora en la reducción de los valores residuales del modelo de RLM. A continuación, se visualizará como quedaría en la base de datos:

	EDAD	SEXO	PROVINCIA	ECIVIL	Tiempo_Vehiculo	Valor_Total_Vehiculo	Prima_Mensual	class	Prediccion
1	54	1	13	1	2	18000.00	64.20	2	63.64533
2	23	2	10	5	2	19999.00	79.28	2	70.17672
3	43	2	19	1	2	38090.00	134.79	1	125.57869
4	55	2	19	5	2	22690.00	86.10	2	77.22777
5	39	1	19	5	3	24999.00	93.54	2	82.70084
7	64	2	19	1	2	29999.00	95.71	1	100.76440
8	22	2	19	5	2	40000.00	113.45	1	131.43647
9	40	2	19	1	2	26999.00	88.08	2	90.44305
10	63	2	19	1	3	26790.00	85.96	2	88.19366
11	40	2	19	1	2	29999.00	84.44	2	99.64375
12	31	1	19	1	2	27000.01	84.04	1	91.56679
13	28	2	19	1	2	29999.00	100.24	1	100.76440
14	33	2	19	5	2	40110.00	116.72	1	131.77383
15	38	2	19	1	2	32000.00	99.88	1	106.90126
16	40	1	19	5	2	17299.00	58.68	2	60.69411
17	46	2	19	1	2	19989.88	77.12	2	68.94677
18	45	1	19	5	2	17388.60	68.04	1	62.08955
19	29	2	19	5	2	17299.00	66.45	1	61.81475
20	44	2	19	1	3	32400.00	123.05	1	106.51962
21	48	2	19	1	2	23990.00	88.20	2	81.21474

Gráfico 17. Valores predictivos. Elaboración propia.

A primeros rasgos, parece ser que la predicción del modelo fue bastante acertada, esto debido a que la diferencia entre la variable dependiente y la variable de respuesta es muy baja. Sin embargo, para poder determinar que tanta discrepancia hay entre estas dos variables, se utiliza la raíz del error cuadrático medio, que permitirá medir esta diferencia y así determinar si lo realizado por el modelo es acertado. Por ende, se requiere del siguiente código:

```
Seguros [Entrenamiento,"RMSE"] <- sqrt (mean(Seguros$Prima_Mensual-
Seguros$Prediccion) ^2)

Seguros[-Entrenamiento,"RMSE"] <- sqrt(mean(Seguros$Prima_Mensual-
Seguros$Prediccion) ^2)
```

En el análisis del modelo reducido, se calculó la Raíz Error Cuadrático Medio (RMSE) tanto para el conjunto de entrenamiento como para el conjunto de prueba. Para llevar a cabo esto, se calcula utilizan dos variables, que son “Prima_Mensual” y “Predicción”, las cuales se eleva al cuadrado la diferencia entre ambas y posteriormente se saca el promedio de todo el conjunto de datos, para finalmente obtener la raíz cuadrada del resultado, concluyendo así la operación matemática. Los resultados de cada uno de los datos se puede evidenciar en la columna creada, que lleva de nombre RMSE.

	EDAD	SEXO	PROVINCIA	ECIVIL	Tiempo_Vehiculo	Valor_Total_Vehiculo	Prima_Mensual	class	Prediccion	RMSE
1	54	1	13	1	2	18000.00	64.20	2	63.64533	0.1182173
2	23	2	10	5	2	19999.00	79.28	2	70.17672	0.1182173
3	43	2	19	1	2	38090.00	134.79	1	125.57869	0.1182173
4	55	2	19	5	2	22690.00	86.10	2	77.22777	0.1182173
5	39	1	19	5	3	24999.00	93.54	2	82.70084	0.1182173
7	64	2	19	1	2	29999.00	95.71	1	100.76440	0.1182173
8	22	2	19	5	2	40000.00	113.45	1	131.43647	0.1182173
9	40	2	19	1	2	26999.00	88.08	2	90.44305	0.1182173
10	63	2	19	1	3	26790.00	85.96	2	88.19366	0.1182173
11	40	2	19	1	2	29999.00	84.44	2	99.64375	0.1182173
12	31	1	19	1	2	27000.01	84.04	1	91.56679	0.1182173
13	28	2	19	1	2	29999.00	100.24	1	100.76440	0.1182173
14	33	2	19	5	2	40110.00	116.72	1	131.77383	0.1182173
15	38	2	19	1	2	32000.00	99.88	1	106.90126	0.1182173
16	40	1	19	5	2	17299.00	58.68	2	60.69411	0.1182173
17	46	2	19	1	2	19989.88	77.12	2	68.94677	0.1182173
18	45	1	19	5	2	17388.60	68.04	1	62.08955	0.1182173
19	29	2	19	5	2	17299.00	66.45	1	61.81475	0.1182173

Gráfico 18. RMSE. Elaboración propia.

Como se puede observar en el gráfico numero 18, el RMSE es muy bajo por ende quiere decir que la predicción es muy acertada y es caspi el mismo valor que el valor de la variable y. Por ende, esto determina que el modelo funciona adecuadamente tanto en los datos de entrenamiento y de prueba.

Continuando con la ejecución del modelo, requerimos de establecer los límites tanto superior como inferior, que puede llegar a tener los valores reales y los valores predichos.

Esto determina, hasta que valor puede llegar a ser el RSME para que la predicción sea acertada. Para esto utilizaremos la siguiente línea de código:

Seguros$L.Inf. <-Seguros$Prediccion- Seguros$RMSE

Seguros$L.Sup. <-Seguros$Prediccion+ Seguros$RMSE

A partir del RMSE calculado, se generaron intervalos de predicción para cada observación en el data frame Seguros. Estos intervalos, que llevan por nombre las columnas L.Inf. (límite inferior) y L.Sup. (límite superior), las cuales se calcularon de la diferencia y sumatoria del valor del RMSE a las predicciones del modelo (Prediccion), respectivamente. Estos intervalos indican un rango dentro del cual se espera que se encuentren los valores reales y predichos de la prima mensual, considerando el error cuadrático medio del modelo:

	EDAD	SEXO	PROVINCIA	ECIVIL	Tiempo_Vehiculo	Valor_Total_Vehiculo	Prima_Mensual	class	Prediccion	RMSE	L.Inf.	L.Sup.
1	54	1	13	1	2	18000.00	64.20	2	63.64533	0.1182173	63.52711	63.76354
2	23	2	10	5	2	19999.00	79.28	2	70.17672	0.1182173	70.05851	70.29494
3	43	2	19	1	2	38090.00	134.79	1	125.57869	0.1182173	125.46047	125.69691
4	55	2	19	5	2	22690.00	86.10	2	77.22777	0.1182173	77.10955	77.34599
5	39	1	19	5	3	24999.00	93.54	2	82.70084	0.1182173	82.58262	82.81906
7	64	2	19	1	2	29999.00	95.71	1	100.76440	0.1182173	100.64618	100.88261
8	22	2	19	5	2	40000.00	113.45	1	131.43647	0.1182173	131.31826	131.55469
9	40	2	19	1	2	26999.00	88.08	2	90.44305	0.1182173	90.32483	90.56127
10	63	2	19	1	3	26790.00	85.96	2	88.19366	0.1182173	88.07544	88.31188
11	40	2	19	1	2	29999.00	84.44	2	99.64375	0.1182173	99.52553	99.76197
12	31	1	19	1	2	27000.01	84.04	1	91.56679	0.1182173	91.44857	91.68501
13	28	2	19	1	2	29999.00	100.24	1	100.76440	0.1182173	100.64618	100.88261
14	33	2	19	5	2	40110.00	116.72	1	131.77383	0.1182173	131.65562	131.89205
15	38	2	19	1	2	32000.00	99.88	1	106.90126	0.1182173	106.78305	107.01948
16	40	1	19	5	2	17299.00	58.68	2	60.69411	0.1182173	60.57589	60.81232
17	46	2	19	1	2	19989.88	77.12	2	68.94677	0.1182173	68.82855	69.06499
18	45	1	19	5	2	17388.60	68.04	1	62.08955	0.1182173	61.97133	62.20776
19	29	2	19	5	2	17299.00	66.45	1	61.81475	0.1182173	61.69653	61.93297

Gráfico 19. Límites superior e inferior. Elaboración propia.

Por tanto, a través de la línea de código anterior se obtuvo los límites tanto inferior como superior, donde se puede observar que los limites son muy estrechos permitiendo así que sea muy fácil de identificar valores fuera de rango. Sin embargo, la mayoría de los datos

están dentro de los límites establecidos, dando por hecho que existe muy poca incertidumbre en los datos, y que las predicciones son completamente adecuadas y robustas para el modelo.

Finalmente, se requiere del gráfico de regresión lineal, que permitirá determinar la linealidad en los datos y cercanía entre los datos, asegurando la convergencia y semejanza entre las variables tanto la dependiente como la variable de repuesta o predicción. Para esto se hará uso de la siguiente línea de código:

```
par (mfrow=c (1,1))

plot(Seguros$Prima_Mensual,Seguros$Prediccion)

abline(0,1)
```

A través del código establecido anteriormente, se arma el ambiente y la gráfica para el modelo de regresión lineal. Se utilizaron las variables "Prima_Mensual" y "Predicción", para visualizar sus datos a través de puntos y así determinar que los principios de la linealidad de la regresión se cumplan.

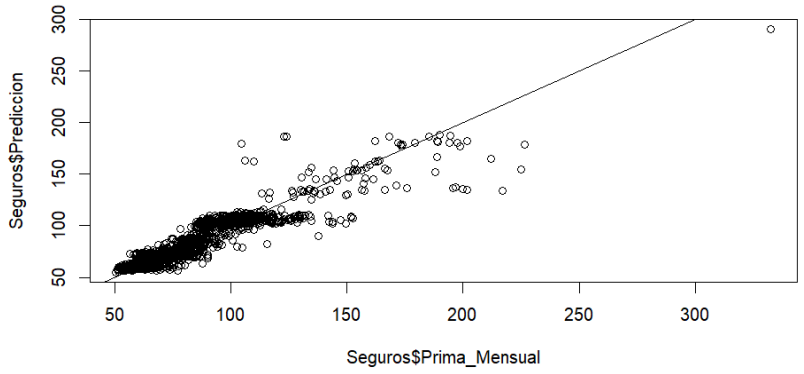


Gráfico 20. Regresión lineal. Elaboración propia.

El gráfico evidencia que el modelo predice de manera precisa para la mayoría de los datos, especialmente en los rangos bajos y medios de la prima mensual, donde las predicciones están cerca de los valores reales. Sin embargo, se observan discrepancias en los valores más elevados, lo que podría deberse a características específicas de estas observaciones o a una falta de variables relevantes para capturar su comportamiento. A pesar de estas desviaciones, el modelo demuestra un buen desempeño general, como lo respalda el bajo RMSE.

Por último, se realizaron dos boxplot que permitieron visualizar el que tan bien se encontraba la estructura de los datos, y si estos poseían pocos valores outliers. Además, el diagrama de caja será de gran de ayuda para identificar si la mediana se encuentra completamente en el centro, dando a entender que los datos están distribuidos de manera simétrica, garantizando que existe normalidad dentro del conjunto de datos. Para esto, se utilizó las siguientes líneas de código, a continuación:

```
Seguros$class <- factor(Seguros$class)

BOX_KMEANS <- ggplot(Seguros, aes(x=Seguros$class,

y=Seguros$Prima_Mensual))+

geom_boxplot(fill="red",color="Black")

BOX_KMEANS
```

En este código podemos observar que primero se transformó en factor la variable “class”, esto debido a que se requiere convertirla de numérica a categórica, para poder clasificar adecuadamente cada boxplot con su respectiva clasificación. Posteriormente, se utilizó “geom_boxplot(fill="red",color="Black"” para poder crear y ejecutar el respectivo

diagrama de bigote, donde su utilizó a la variable “class” como independiente, y a la variable “Prima_Mensual” como dependiente, todo esto en el ambiente de “ggplot”. Para concluir, se utilizó “BOX_KMEANS” para visualizar el gráfico a continuación.

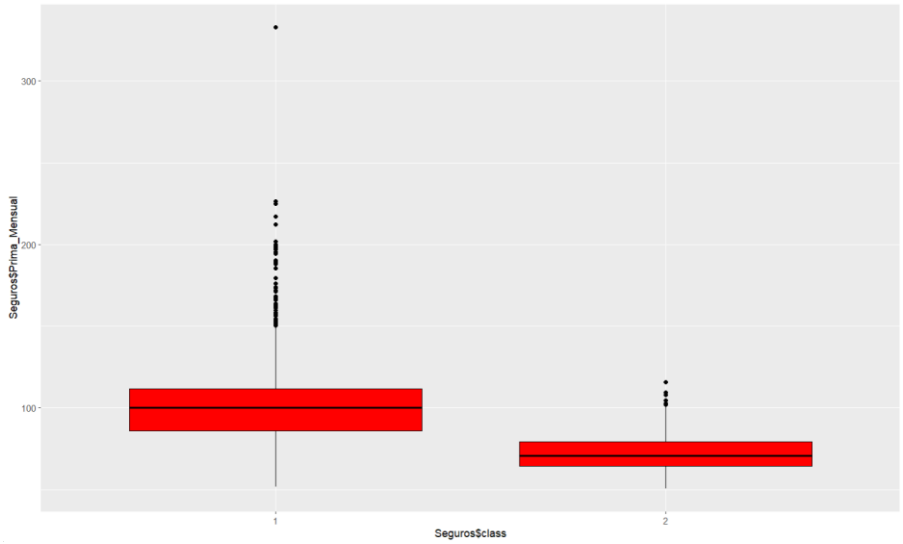


Gráfico 21. Boxplot de Segmentación. Elaboración propia.

En este gráfico, se observa que en el primer boxplot, la mediana se encuentra cerca del 100, por ende, quiere decir que la gran mayoría de primas mensuales de la “class 1” oscilas entre los $100 mensuales que paga el cliente, mientras que en el segundo boxplot, “class 2”, la mediana se encuentra cerca del 60. Por consiguiente, esto sugiriere que la gran mayoría de primas oscilan entre los $60 dólares, dando a entender que el primer diagrama de bigotes sería un segmento más premium mientras que el segundo un segmento mas estándar o económico.

Comparación con artículo similar

Título del trabajo: Factores de Riesgo de Siniestralidad y Cálculo de Primas de los Vehículos Asegurados en el Ecuador mediante Modelos Lineales Generalizados

Autores: Quishpe, I. (2015)

El trabajo utilizado para la comparación se basa en el análisis de la gravedad de una siniestralidad vehicular y su probabilidad de que el evento ocurra; de este modo, logra establecer unos parámetros para que las aseguradoras calculen las primas. Sin duda, este es un caso similar al trabajo presente, ya que ambos buscan desarrollar un modelo basado en la estadística para el cálculo de la prima vehicular.

No obstante, existen ciertas discrepancias entre ambos estudios. Comenzado por el hecho de que, en el otro documento, se utiliza el modelo logit en la aplicación *Stat*, se toma en consideración la siniestralidad como factor clave. Mientras que, en este caso, se hizo uso de la regresión lineal múltiple y K-Means mediante el aplicativo *R-Studio* y no se considera la siniestralidad ni algún factor similar.

En ambos casos, se puede afirmar que el propósito del trabajo es mejorar la competitividad a través de una nueva alternativa para la tarificación de primas vehiculares. En síntesis, ambas investigaciones realzan la importancia del sector segurador dentro del país y la necesidad de seguir implementando nuevas formas para optimizar el cálculo de primas usando herramientas estadísticas.

Conclusiones

El objetivo principal conseguido en este trabajo investigativo radicó en predecir el cálculo de las primas vehiculares y posteriormente, la segmentación de clientes de seguros vehiculares, utilizando técnicas de aprendizaje automatizado o también conocido como Machine Learning. Para aquello, se desarrolló un modelo predictivo, usando los algoritmos

de RLM y K means, lo que permitió obtener la meta esperada y obtener conclusiones significativas del estudio.

Mediante la revisión de la literatura, se pudo concluir que los conceptos de machine learning son altamente beneficiosos, con respecto a este trabajo investigativo. Se encontraron diversos estudios que establecen que los modelos de aprendizaje automatizado facilitan la optimización de procesos, la reducción de flujos de trabajo y finalmente, ofrecen una predicción muy precisa en contexto de valores inciertos que buscamos conseguir.

El respectivo análisis de datos del modelo permitió identificar y obtener que existen dos categorizaciones o segmentaciones de clientes. Estas clases, se pueden deber a que un grupo tiene vehículos menos nuevos y con primas más baratas y el segundo grupo corresponder a que son vehículos más nuevos con primas más caras. Así, se puede determinar que la utilización de este nuevo modelo de cálculo es sustancial, ya que se podrá fijar tarifas en base al contexto económico del cliente, permitiendo así que la empresa tenga mayor conocimiento y respaldo de sus clientes en el mercado.

La implementación y ejecución del modelo predictivo de aprendizaje automatizado, fue completamente un éxito, de los valores significantes que afectaban un poco el mismo. El modelo puedo predecir el valor de la prima vehicular con una Raíz del Error Cuadrático Medio (RMSE) muy bajo. Permitiendo así comprender, que no existe mayor diferencia entre lo real y lo predicho, validando así la eficacia del modelo y robustez de este.

Gráficamente, se pudo establecer que, en base al modelo ejecutado, las variables utilizadas poseen una gran correlación. Permitiendo así, que futuramente si se utilizan mayor cantidad de variables tanto sociodemográficas como del vehículo, estas incluso podrán

otorgar una mayor precisión de cuanto debería pagar exactamente una persona por la prima de su vehículo. Sin embargo, es importante también entender que siempre se debe medir la participación de las variables dentro del modelo y si realmente están formando parte de la predicción, o netamente forman parte de esos valores outliers que deben ser eliminados.

La utilización de estos modelos facilita la identificación y generación continua de oportunidades de mejora en el cálculo de las primas vehiculares, que permitirá una mayor satisfacción y justicia a los clientes al momento de tener que pagar su prima de seguro de vehículo. Generando así, una mayor confianza de estos con los seguros vehiculares y prometiendo un aumento prospero en las ventas futuras de este producto de seguro.

Referencias

¿Qué es el agrupamiento en clústeres? (s. f.). Google For Developers. https://developers.google.com/machine-learning/clustering/overview?hl=es-419

¿Qué es el machine learning? - Explicación sobre el machine learning empresarial - AWS. (s. f.-c). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/machine-learning/

¿Qué es la prima de un seguro? (2024). Banco Santander. https://www.bancosantander.es/glosario/prima-seguro

¿Qué es un seguro vehicular? Su importancia y más. (2021). Diners Club. https://www.dinersclub.com.ec/experiencias/diners-club/importancia-seguro-vehicular

¿Todo lo que debes saber sobre una póliza de seguro? (2024). Seguros del Pichincha. https://segurosdelpichincha.com/blogs/poliza-seguro-condiciones-coberturas-ecuador

adSalsa, E. (2024, 8 marzo). Segmentación de audiencia: más allá con IA. adSalsa. https://www.adsalsa.com/segmentacion-de-audiencia-con-ia/#:~:text=Utilizar%20la%20IA%20para%20segmentar,sus%20datos%20de%20contacto%2C%20realizar

adSalsa, E. (2024, 8 marzo). Segmentación de audiencia: más allá con IA. adSalsa. https://www.adsalsa.com/segmentacion-de-audiencia-con-ia/#:~:text=Utilizar%20la%20IA%20para%20segmentar,sus%20datos%20de%20contacto%2C%20realizar

Anwar, M. (2024, 8 marzo). What is Data Cleansing? A Complete Guide | Astera. *Astera*. https://www.astera.com/es/type/blog/data-cleansing/

Asamblea Nacional del Ecuador. (2008). Constitución de la República del Ecuador. Actualización enero de 2021. Recuperado de https://www.defensa.gob.ec/wp-content/uploads/downloads/2021/02/Constitucion-de-la-Republica-del-Ecuador_act_ene-2021.pdf

BBVA ESPAÑA & BBVA. (2024, 31 octubre). ¿Qué diferencia hay entre los seguros de vida y no vida? BBVA. https://www.bbva.es/finanzas-vistazo/ef/seguros/diferencias-entre-el-seguro-de-vida-y-no-vida.html#:~:text=Tambi%C3%A9n%20hay%20seguros%20de%20vida,caso%20de%20los%20seguros%20m%C3%A9dicos).

Boden, M. (2016). Inteligencia Artificial. Turner Publicaciones.

Bonta, P., Farber, M. (1995). 199 Preguntas Sobre Marketing y Publicidad. Norma.

Cajamarca, I. (2022, 23 septiembre). Los países de América Latina que tienen los seguros para automotores más costosos. Diario la República. https://www.larepublica.co/globoeconomia/los-paises-de-america-latina-que-tienen-los-seguros-para-automotores-mas-costosos-3453959

Carrasquilla-Batista, A., Chacón-Rodríguez, A., Núñez-Montero, K., Gómez-Espinoza, O., Valverde, J., & Guerrero-Barrantes, M. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Revista Tecnología en Marcha*, *29*, 33-45.

Chen, H., Chen, G., Peng, E. (2013). Business Intelligence and Analytics: Research Directions. ACM Digital Library. Association for Computing Machinery, Estados Unidos, Nueva York. https://doi.org/10.1145/2407740.2407741

Ciencia de Datos. (s. f.). Regresión lineal múltiple. Recuperado de https://cienciadedatos.net/documentos/25_regresion_lineal_multiple

Colomé, J. (2012). Aproximación a la lógica difusa: una aplicación a la valoración de activos. Universitat Oberta de Catalunya. https://openaccess.uoc.edu/bitstream/10609/30901/1/COLOME_WP2012_Aproximacion.pdf

Daniel. (2023, 30 octubre). Machine Learning: definición, funcionamiento, usos. Formación En Ciencia de Datos | DataScientest.com. https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

De los Santos, P. R. (2023, 29 junio). Datos de entrenamiento vs datos de test. Telefónica Tech. https://telefonicatech.com/blog/datos-entrenamiento-vs-datos-de-test

Fanjul, J. M. (2024, 4 junio). ¿Qué es el Análisis de Componentes Principales y cómo reducir el tamaño de una base de datos? Blog de Hiberus. https://www.hiberus.com/crecemos-contigo/analisis-de-componentes-principales/

Fedeseg. (2022, 14 septiembre). Cifras que dejó la pandemia en el ramo de Vehículos y expectativas para el 2022. Fedeseg1. https://www.fedeseg.org/post/cifras-que-dej%C3%B3-la-pandemia-en-el-ramo-de-veh%C3%ADculos-y-expectativas-para-el-2022

Fernández, O. (2024). RStudio: Simplifica tu análisis de datos y el cálculo estadístico. https://aprenderbigdata.com/rstudio/

Gobierno del Ecuador. (2023). Decreto Ejecutivo No. 904. Recuperado de https://www.telecomunicaciones.gob.ec/wp-content/uploads/2023/11/Decreto-Ejecutivo-No.-904.pdf

Gonzalez, J. L. (2020, 13 julio). Tipos de aprendizaje automático - SoldAI - Medium. Medium. https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2

Gonzalez, L. (2022, 7 septiembre). Regresión polinomial – teoría. Aprende IA. https://aprendeia.com/algoritmo-regresion-polinomial-machine-learning/

GraphEverywhere, E. (2019, 29 octubre). Algoritmo de distancia euclidiana. GraphEverywhere. https://www.grapheverywhere.com/algoritmo-de-distancia-euclidiana/

GraphEverywhere, E. (2019, octubre 29). Algoritmo de similitud de coseno. GraphEverywhere. https://www.grapheverywhere.com/algoritmo-de-similitud-de-coseno/

Hill, C., Jones, G. (2011). Administración Estratégica. McGRAW-HILL/INTERAMERICANA EDITORES, S.A. de C.V

IBM Cognos Analytics 11.1.x. (s. f.). https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-analysis-variance-anova

Ibm. (2024, 13 mayo). ¿Qué es ETL (extraer, transformar, cargar)? | IBM. IBM. https://www.ibm.com/es-es/topics/etl

Ibm. (2024, septiembre 12). Clustering. IBM. https://www.ibm.com/es-es/topics/clustering#:~:text=%C2%BFQu%C3%A9%20es%20el%20clustering%3F&text=IBM,basados%20en%20similitudes%20o%20patrones.

inbestMe. (2023, 9 julio). ¿Qué es la correlación? inbestMe ES. https://www.inbestme.com/es/es/blog/que-es-la-correlacion/

Jarroba, R. [. (2017, 2 junio). Selección del número óptimo de Clusters - Jarroba. Jarroba. https://jarroba.com/seleccion-del-numero-optimo-clusters/

K, R. (2024, 30 enero). The Role of Machine Learning in Insurance Pricing Models. Medium. https://medium.com/@raghav_17189/the-role-of-machine-learning-in-insurance-pricing-models-cd9a382c0be4#:~:text=Machine%20Learning%20in%20Insurance%20Pricing%20Model%3A%20The%20Impact&text=Machine%20learning%20can%20analyze%20vast,understanding%20of%20various%20risk%20factors.

Learn Statistics Easily. (2024, 24 julio). Qué es: Distancia de Manhattan - APRENDE ESTADÍSTICAS FÁCILMENTE. LEARN STATISTICS EASILY. https://es.statisticseasily.com/glossario/what-is-manhattan-distance/

Learn Statistics Easily. (2024, 28 septiembre). *Qué es: Convergencia K-Means explicada*. LEARN STATISTICS EASILY. https://es.statisticseasily.com/glosario/%C2%BFQu%C3%A9-significa-k-convergencia%3F/

Learn Statistics Easily. (2024, julio 24). Qué es: Error absoluto medio (MAE): APRENDA ESTADÍSTICAS FÁCILMENTE. LEARN STATISTICS EASILY. https://es.statisticseasily.com/glosario/%C2%BFQu%C3%A9-significa-error-absoluto-mae%3F/

Lilian Judith Sandoval. (2018). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS [PDF]. Recuperado de http://redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

López, J. F. (2024, 1 agosto). ¿Qué es el coeficiente de determinación? Cálculo y ejemplos. Economipedia. https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html

Machine Learning in Insurance: Applications, Use Cases, and Projects. (2024, 28 octubre). ProjectPro. https://www.projectpro.io/article/machine-learning-in-insurance/774#:~:text=Additionally%2C%20machine%20learning%20can%20consider,precise%20prediction%20than%20standard%20approaches.

Machine learning: Definición, Ventajas y Desventajas. (s. f.). Salesforce. https://www.salesforce.com/es/resources/definition/machine-learning/#topic4

Marrero, L., Carrizo, D., García-Santander, L., & Ulloa-Vásquez, F. (2021). Uso de algoritmo K-means para clasificar perfiles de clientes con datos de medidores inteligentes de consumo eléctrico: Un caso de estudio. *Ingeniare. Revista chilena de ingeniería*, *29*(4), 778-787.

Navarro, S. (2024, 16 abril). ¿Qué es el clúster por densidad? | KeepCoding Bootcamps. KeepCoding Bootcamps. https://keepcoding.io/blog/que-es-el-cluster-por-densidad/

Negash, S., Gray, P. (2008). Business Intelligence. In: Handbook on Decision Support Systems 2. International Handbooks Information System. Springer, Berlín, Heidelberg. https://doi.org/10.1007/978-3-540-48716-6_9

Oracle® Fusion Cloud EPM Trabajo con Planning. (s. f.-b). https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_RMSE.htm#PFUSU-GUID-FD9381A1-81E1-4F6D-8EC4-82A6CE2A6E74

Oracle® Fusion Cloud EPM Trabajo con Planning. (s. f.). https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_RMSE.htm

Pérez, J., Gardey, A. (2021). Variable - Qué es, usos, ejemplos, tipos y en la informática. https://definicion.de/variable/

Prateek. (2022, 20 marzo). K means clustering algorithm. KeyToDataScience. https://keytodatascience.com/k-means-clustering-algorithm/

Redacción. (2019, 25 febrero). Los orígenes del seguro de autos. La Vanguardia. https://www.lavanguardia.com/seguros/coches/20190225/462107344505/los-origenes-del-seguro-de-autos.html

República del Ecuador. (2021). Ley Orgánica de Protección de Datos Personales. Recuperado de https://www.finanzaspopulares.gob.ec/wp-content/uploads/2021/07/ley_organica_de_proteccion_de_datos_personales.pdf

ResearchGate. (s.f.). Partitional Clustering [Figura]. ResearchGate. https://www.researchgate.net/figure/Partitional-Clustering_fig2_312590567

Rodríguez, D. (2023, 11 junio). Número óptimo de clústeres con Silhouette e implementación en Python. Analytics Lane. https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhouette-e-implementacion-en-python/

Roldán, P. N. (2022, 24 noviembre). Modelo de regresión - Definición, qué es y concepto | Economipedia. Economipedia. https://economipedia.com/definiciones/modelo-de-regresion.html

Rouhiainen, L. (2018). Inteligencia Artificial. 101 Cosas Que Debes Saber Hoy Sobre Nuestro Futuro. Editorial Alienta.

RPubs - Final proyect (Chap 4). (s. f.). https://rpubs.com/Margerithe/1043851

RPubs - K means. (s. f.). https://rpubs.com/JosueEmmanuel/Kmeans

RStudio Desktop - Posit. (2024, 18 julio). Posit. https://posit.co/download/rstudio-desktop/

Rueda, J. F. V. (2019, 9 agosto). Aprendizaje supervisado y no supervisado - healthdataminer.com. healthdataminer.com. https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/#:~:text=Los%20m%C3%A9todos%20no%20supervisados%20(unsupervised,ya%20sea%20categ%C3%B3rico%20o%20num%C3%A9rico.

Saavedra, J. A. (2023, 3 mayo). Regresión Lineal: teoría y ejemplos. Ebac. https://ebac.mx/blog/regreson-lineal

San Sebastián, C. (s. f.). Tema 14: Clustering. [PDF]. Recuperado de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t14clustering.pdf

Sancho, F. (s.f.). Aprendizaje supervisado y no supervisado. Universidad de Sevilla. https://www.cs.us.es/~fsancho/Blog/posts/Aprendizaje_Supervisado_No_Supervisado.md.html

silhouette_score. (s. f.). Scikit-learn. https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.silhouette_score.html

St Onge, K. J. (2008, 27 febrero). First auto policy sold 110 years ago today. Insurance Journal. https://www.insurancejournal.com/news/national/2008/02/27/87696.htm

Statista. (2024, 16 julio). Seguros de no vida: ingresos mundiales por emisión de primas por segmento 2017-2028. https://es.statista.com/estadisticas/1479209/volumen-de-primas-seguros-de-no-vida/#statisticContainer

Statute Law Database. (s. f.). Road Traffic Act 1930. https://www.legislation.gov.uk/ukpga/Geo5/20-21/43#:~:text=An%20Act%20to%20make%20provision,protection%20to%20amend%20the%20Assurance

Tatvasoft. (2023, 7 agosto). ETL Process (Extract Transform Load). TatvaSoft Blog. https://www.tatvasoft.com/blog/etl-process-extract-transform-load/

The Black Box Lab. (2022, 30 junio). Machine Learning: Algoritmos de clasificación y regresión - The Black Box Lab. https://theblackboxlab.com/2022/05/06/machine-learning-diferencias-entre-algoritmos-clasificacion-regresion/

Tipos de clustering jerárquico. (s. f.). AI Planet (Formerly DPhi). https://aiplanet.com/notebooks/4381/felipe.cortes/tipos-de-clustering-jerarquico

Torrealba, D. F. (2016, 19 agosto). Historia y leyenda del primer seguro de coche. elEconomista.es. https://www.eleconomista.es/ecomotor/motor/noticias/7770897/08/16/Historia-y-leyenda-del-primer-seguro-de-coche.html

Universidad de Santiago de Compostela. (s.f.). *Regresión lineal múltiple: El modelo, estimación de los parámetros, contrastes* [PDF]. Recuperado de http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140128_RegresionMultiple.pdf

Zambrano, R. (2019). Qué es R y por qué utilizarlo. https://openwebinars.net/blog/que-es-r-y-por-que-utilizarlo/

DECLARACIÓN Y AUTORIZACIÓN

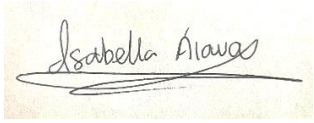
Nosotros, **Álava Llusca, Isabella Dominique** con C.C: # 0925828675 y **Gordon Sánchez, Flavio Paúl** con C.C: # 0954223202 autores del trabajo de integración curricular: **Modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación del consumidor,** previo a la obtención del título de **Licenciado en Negocios Internacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de integración curricular para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de integración curricular, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

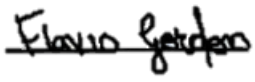
Guayaquil, a los 07 del mes de febrero del año 2025

LOS AUTORES:



f. _____

Álava Llusca, Isabella Dominique
C.C: # 0925828675



f. _____

Gordon Sánchez, Flavio Paúl

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE INTEGRACIÓN CURRICULAR			
TEMA Y SUBTEMA:	Modelo de aprendizaje automatizado para el cálculo de la prima vehicular y segmentación del consumidor.		
AUTOR(ES)	Álava Llusca, Isabella Dominique Gordon Sánchez, Flavio Paúl		
REVISOR(ES)/TUTOR(ES)	Ing. Carrera Buri, Félix Miguel, Mgs.		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	Facultad de Economía y Empresa		
CARRERA:	Negocios Internacionales		
TITULO OBTENIDO:	Licenciado en Negocios Internacionales		
FECHA DE PUBLICACIÓN:	07 de febrero de 2025	**No. DE PÁGINAS:**	98
ÁREAS TEMÁTICAS:	Marketing, experiencia del cliente		
PALABRAS CLAVES/ KEYWORDS:	Industria seguros vehiculares, machine learning, prima vehicular.		

RESUMEN/ABSTRACT:

El presente trabajo investigativo utiliza los conceptos y algoritmos de Machine Learning, Regresión Lineal Múltiple y K-Means, para mejorar la precisión en el cálculo de las primas vehiculares y la segmentación de sus respectivos clientes. Basado en esto, se desarrolló un modelo de Aprendizaje Automatizado, que predice a través de variables como el sexo, valor vehículo, valor de prima actual, entre otras, cuanto sería la prima por pagar de cada cliente al momento de asegurar el vehículo. Además, con estas mismas variables el modelo también segmenta a los clientes dependiendo de sus características, dando así una mayor personalización de las primas vehiculares y determinado las categorías de posibilidad a pagar en los clientes, permitiendo así que futuros clientes que puedan tener características similares, se pueda tener una idea clara de cuanto deben de pagar por el valor de la prima vehicular.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:**	E-mail: isabella.alava@cu.ucsg.edu.ec flavio.gordon@cu.ucsg.edu.ec
CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UIC):	**Nombre: Freire Quintero Cesar enrique**	
	Teléfono: +593-990090702	
	E-mail: cesar.freire@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA	
Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	