Transición hacia un modelo para la Industria de Seguros de Vehículos. 'Paga Según Manejas' – Grupo 5

Resumen

Cada vez un mercado más competitivo y complejo como el sector de seguros requiere una mayor innovación para mantener y mejorar cada vez más su participación en el mercado; por lo que se propone un nuevo método denominado "paga según manejas" para la clasificación de asegurados para una compañía de seguros en Bogotá, Colombia, basado en información de siniestralidad pública de la ciudad. El modelo incluye la incorporación de análisis de la información, con características categóricas, ordinales y numéricas comprendiendo desafíos adicionales a los modelos supervisados presentes en la actualidad. Propendiendo en superar el modelo actual de categorías de tarifas de seguros de automóviles, que depende principalmente de las características de los vehículos y sus propietarios lo que permitirá unas categorías de tarifas más precisas y justas que beneficien el buen comportamiento de los conductores mientras siguen siendo rentables y sostenibles para la compañía de acuerdo con las posibles anomalías presentes en las reclamaciones. Nuestra contribución clave será la identificación de patrones en accidentes viales y la categorización de asegurados según la información de accidentes del periodo. Para este desarrollo se implementan técnicas analíticas de clasificación de aprendizaje no supervisado, estimaciones de métricas y medidas de similitud de los accidentes analizados.

Introducción

En un entorno cada vez más retador, la industria de seguros en Bogotá y el mundo se enfrenta al desafío de adaptarse a las necesidades cambiantes de los conductores y mejorar la equidad y personalización en las tarifas de seguros; según las estadísticas proporcionadas por Fasecolda (2023) se expiden al año alrededor de 7.8 millones de pólizas SOAT y se realizan pagos en Colombia por cerca de 1.600 millones de pesos anuales en siniestros viales los cuales se vienen incrementando en tasas de hasta el 12% anual.

En este contexto, surge la pregunta: ¿Cómo podemos ofrecer un sistema de seguros más justo y preciso que refleje verdaderamente el comportamiento de conducción de los asegurados?; que permita a la compañía de seguros mantener su liderazgo y solidez en el mercado al ofrecer un enfoque innovador y equitativo para determinar las tarifas de seguros.

La literatura sobre modelos de aplicación en el sector de seguros de automóviles sin supervisión es extremadamente escasa. Con las consultas realizadas se encuentran aplicaciones realizadas especialmente en detecciones de fraudes empleando modelos supervisados, métodos de clasificación espectral no supervisado para anomalías (SRA), SVM, análisis de componentes principales y agrupación como k-means y agrupamiento jerárquico entrenados principalmente con características numéricas; desde donde surte la idea de la presentación de modelos innovadores, por lo que la propuesta es incluir valores categóricos y ordinales que ayuden en la identificación de similitud en los siniestros ocurridos.

El contexto organizacional es el mercado de seguros en Bogotá, caracterizado por una competencia feroz y la necesidad de diferenciarse en un mercado saturado.

Se desarrolla un modelo de aprendizaje no supervisado, aprovechando técnicas avanzadas de análisis de datos para abordar la problemática; y poder responder la pregunta fundamental del negocio de cómo podemos utilizar datos de siniestralidad y comportamiento de conducción para categorizar a los asegurados de manera más precisa y detectar posibles casos de fraude en accidentes.

Revisión preliminar de la literatura

La búsqueda por generar una mejor discriminación de tarifas para el mercado de seguros de vehículos ha conllevado al desarrollo de metodologías que logren establecer tarifas en función de características de las personas al manejar, principalmente en el contexto internacional, en donde se han implementado diferentes estudios centrados en países desarrollados.

De acuerdo con el estudio Reese y Pash (2009) para poder establecer un sistema de tarifas diferenciadas optimo es necesario partir de la implementación de incentivos económicos a los usuarios para que estos permitan capturar información de su estilo de conducción, ya sea por medio de teléfonos inteligentes o GPS, y así, evaluar mejores tarifas, por medio de recolectar información de la distancia recorrida, los tipos de carretera en los que transita, si conduce en horas de mayor riesgo de accidentes, y su frecuencia de viajes.

Entre los estudios que aplican el concepto de “paga como conduces” se encuentra el desarrollado por Carfora, Martinelli, Mercaldo, Nardone, Orlando, Santone y Vaglini (2018) en el cual realizan una clasificación de estilo de conducción de los conductores en términos de agresividad y el riesgo asociado a este estilo.

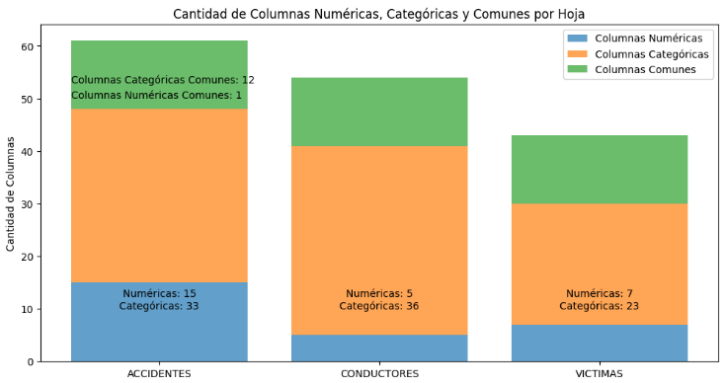
Para esto, haciendo uso de diferentes métricas asociadas a las características de los vehículos y características asociadas con la latitud y longitud por donde transita el conductor, junto con el horario, buscan clasificar en primera instancia el tipo de carretera por donde se encontraba transitando el vehículo a la hora de capturar la información. Esto lo logran haciendo uso del algoritmo k-means, mediante el cual definen 2 clústers asociados a los tipos de carreteras urbanas y de autopista, y tras evaluar la desviación estándar de la aceleración en cada clúster identifican los estilos de conducción más agresivos.

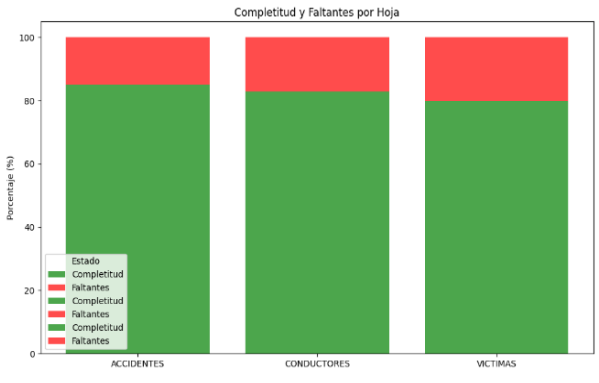
Las investigaciones disponibles sobre el tema se basan en datos recopilados posterior a la adquisición de un seguro. Estos datos se utilizan para crear clasificaciones del público objetivo en función de variables como las rutas, velocidad, por donde conducen y sus horarios, reflejando que aquellos con mejores patrones de conducción y menor exposición a riesgos son quienes son elegibles para mejores tarifas más económicas. En este estudio por medio de la información de siniestralidad de la ciudad de Bogotá, y mediante algoritmos de aprendizaje no supervisado descritos en la propuesta metodológica, se brindará una visión alternativa en la tarificación de seguros, por medio de la identificación de patrones en accidentes viales, su gravedad y la caracterización de los conductores involucrados en estos.

Descripción de los datos

Con el fin de resolver el problema planteado se emplean los datos de https://www.movilidadbogota.gov.co/web/simur correspondiente a la consolidación de siniestros viales ocurridos durante el año 2019 en la ciudad de Bogotá, incluyendo información detallada de cada evento, ubicación geográfica, vehículos involucrados, información de conductores y caracterización de las víctimas.

El dataset original incluye una tabla separada en cada una de las dimensiones mencionadas anteriormente, Accidentes (34990, 50), Conductores (66179, 42), Victimas (9465, 31), en cada uno de ellos se incluyeron variables categóricas y numéricas, como se muestra en el siguiente gráfico:

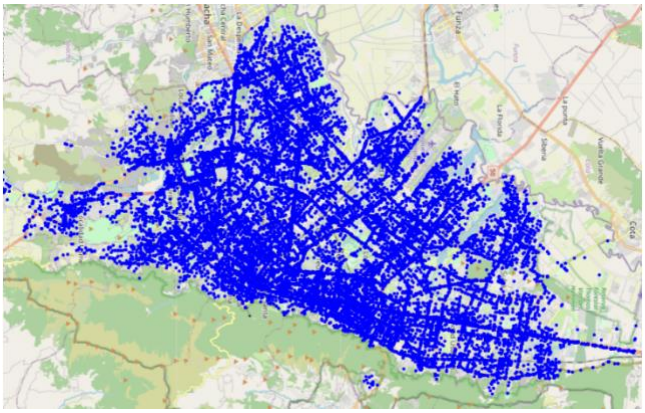


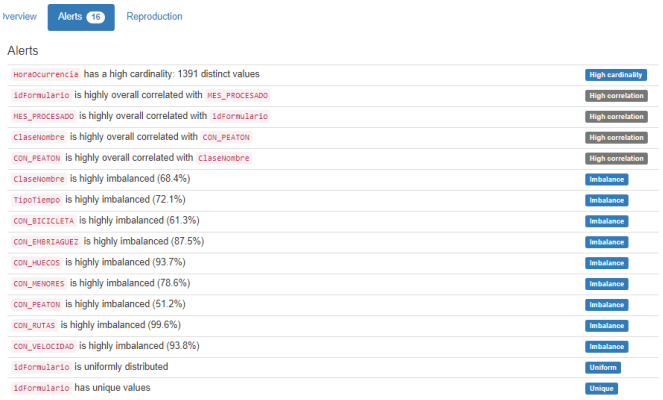


Dentro de los dataset la descripción más relevante de cada uno es:

- Accidentes (34990, 19):

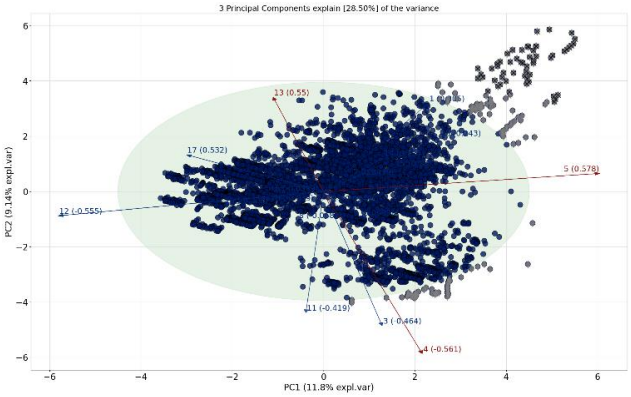
Los datos de accidentes contienen la descripción del tipo de accidente, fecha, ubicación, tipo de vía, y marcaciones si el accidente involucra motos, peatones, personas mayores, exceso de velocidad, entre otros. En el siguiente gráfico se encuentra la distribución por ubicación de los accidentes registrados en la ciudad, marcando especialmente las avenidas principales de la ciudad y la zona oriente. En el preprocesamiento se validó la completitud de las columnas, y para los campos de Latitud y Longitud se imputaron 1.007 registros, de acuerdo con la similitud de la dirección frente a los otros registros. Dentro de los accidentes más comunes se encuentran los choques, los cuales representan el 85% (29.700) de los accidentes. Así mismo, solo el 1,4% corresponde a eventos en donde se presentan fallecimientos.

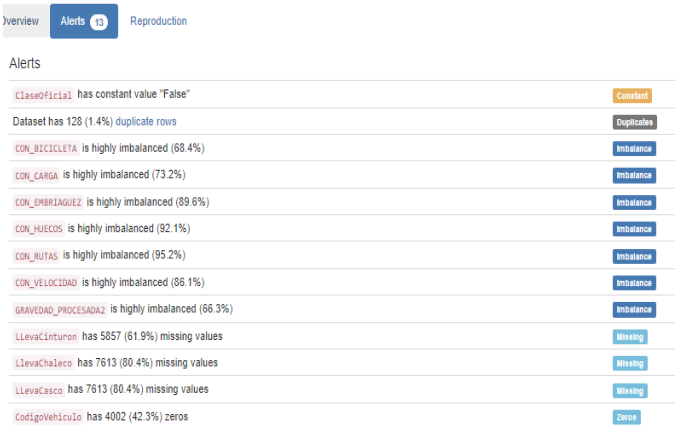




- Victimas (9465, 26):

Los datos procesados permiten identificar características demográficas y comportamientos durante el accidente de acuerdo con las dimensiones presentes en la información para lo cual fue necesario la imputación de valores faltantes con técnicas de imputación múltiple, sencilla y media; transformación de variables categóricas a ordinales para medir la gravedad de la víctima y la conversión de columnas que indican la protección de la víctima (Ej. “LlevaCinturon”), el incremento de gravedad y columnas que inician “CON_” (Ej. “CON_EMBRIAGUEZ). Finalmente, se eliminan todas las columnas con valores vacíos completamente.

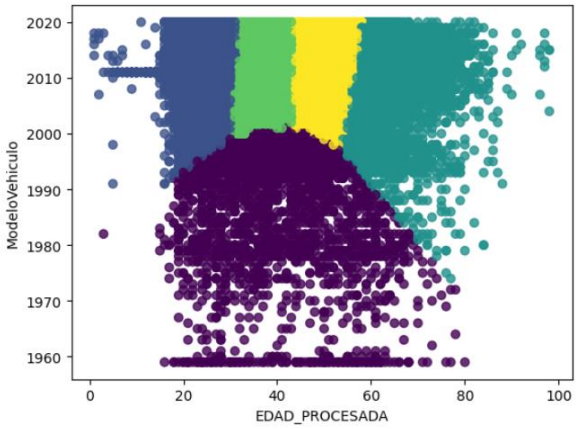




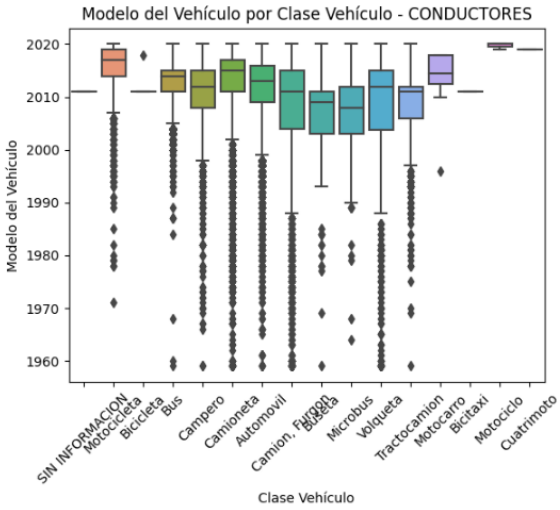
- Conductores (66179, 24):

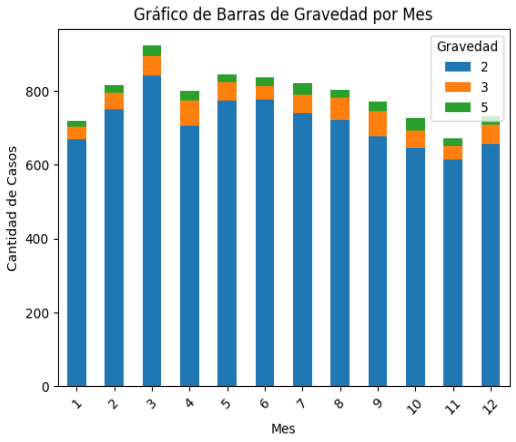
Los datos de conductores contienen características del conductor y del vehículo que conducía en el accidente como: fecha, sexo, licencia y restricciones, vehículo y sus características, cantidad pasajeros,

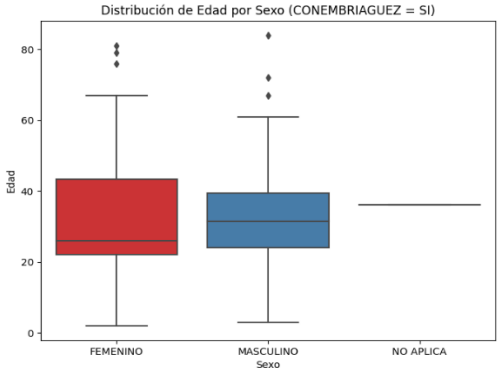
capacidad de carga, si llevaba o no elementos como casco, chaleco o cinturón y gravedad de las heridas, entre otros. Para limpiar la base también se aplicaron técnicas de imputación de datos, eliminación de columnas en su mayoría vacías, y selección de columnas relevantes. En el siguiente gráfico se evidencian 5 grupos de conductores basados en su edad y el modelo del vehículo que conducían en el momento del accidente, adicionalmente se presentan algunos hallazgos importantes en los datos como las correlaciones entre variables, por ejemplo, una interesante es que la gravedad procesada está altamente correlacionada a lleva chaleco y lleva casco:

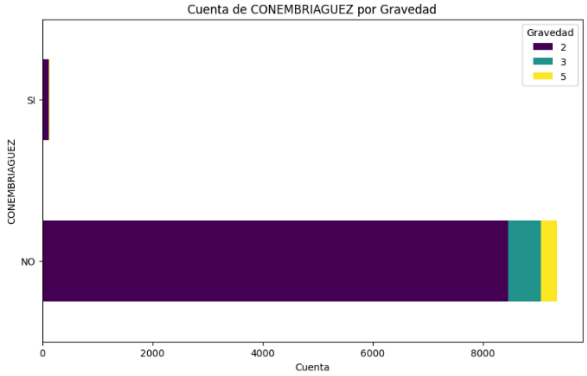


Se realizan algunos análisis de los datos de acuerdo con las características que puedan responder a las preguntas planteadas:









Propuesta metodológica

1. Preparación de los datos:

Para abordar este proyecto, debemos preprocesar y agregar los datos de tres tablas iniciales (Accidentes, Víctimas y Conductores) en dos tablas transformadas (Accidentes y Conductores) que resuman la información relevante presente en otras tablas para trabajar y lograr los dos propósitos específicos que tenemos: caracterizar los conductores según su historial de siniestros e identificar posibles fraudes con base en siniestros anómalos. Para lograrlo entonces realizaremos:

a. La exploración inicial y el preprocesamiento de las diferentes tablas para identificar y seleccionar las columnas de interés y tratar los datos faltantes.
b. Una ETL para crear dos tablas que resuman información de las bases originales:
    i. CONDUCTORES: Esta tabla se caracterizará por almacenar la información original de la tabla de conductores depurada, y, además, incluirá columnas agregadas de otras tablas como por ejemplo cantidad de víctimas graves en accidentes, cantidad de accidentes en el último año, etc. Esto permitirá que cada fila contenga información tanto del conductor como del accidente y las víctimas.
    ii. ACCIDENTES: Esta tabla se caracterizará por almacenar la información original de la tabla de accidentes depurada, y, además, incluirá columnas agregadas de otras tablas como por ejemplo cantidad de víctimas graves en el accidente, cantidad de conductores involucrados, etc. Esto permitirá que cada fila contenga información tanto del accidente como de los conductores y víctimas.

2. Exploración de los datos:

En esta fase exploramos los datos; crucial para identificar patrones en el comportamiento de conducción, segmentar tarifas justas y detectar fraudes en accidentes.

Analizamos los datos de las nuevas tablas transformadas con columnas agregadas de CONDUCTORES y ACCIDENTES para lograr identificar preguntas como: ¿Cuál es el promedio de heridos por accidente?, ¿Qué tan frecuente es que haya un muerto en los accidentes?,¿Cuáles son las características de los conductores en los accidentes donde se presentan víctimas?, entre otras.

3. Luego, procedemos a aplicar técnicas de aprendizaje no supervisado para cada uno de los objetivos perseguidos en este proyecto:

Segmentación de categorías de tarifas según el riesgo del propietario/conductor:

a. Aplicar algoritmos de aprendizaje no supervisado para la segmentación de tarifas como K-Means, K-Medioides, Clústering Jerárquico, DBScan y PCA.
b. Evaluar la calidad de los clústers formados y asignar etiquetas a los conductores según el clúster al que pertenecen.
c. Realizar un análisis de los clústers para determinar si existen diferencias significativas en términos de comportamiento de conducción y riesgo. Para respaldar la segmentación de conductores/propietarios basada en el riesgo percibido.

Detección de anomalías en siniestros:

d. Utilizar algoritmos de detección de anomalías para identificar siniestros que puedan indicar fraude, incluyendo DBScan, Isolation Forest, entre otros.
e. Establecer umbrales para etiquetar siniestros inusuales como posibles casos de fraude.

4. Evaluación y validación:

a. Evaluar la efectividad de la segmentación de categorías de tarifas basadas en los clústers de conductores mediante métricas relevantes como la cohesión, separación entre clústers, y otras técnicas cualitativas que desde un punto de vista de negocio nos ayuden a entender si estos clústers logran clasificar bien el riesgo.
b. Medir la efectividad de los posibles fraudes por medio de un análisis descriptivo de negocio que permita evaluar si efectivamente los siniestros identificados como posible fraude si tienen características que alguien que desea realizar un fraude lo pueda simular. (Esto frente a la

imposibilidad de tener etiquetas para hacer evaluación real en donde podamos tener una variable Y que nos permita calcular métricas más precisas como la precisión o el F1-Score).

Finalmente, dados los clústers encontrados buscamos definir unas categorías asociadas al riesgo que comporta cada clúster y crear un algoritmo supervisado con las etiquetas de los clústers para que la empresa de seguros pueda clasificar los nuevos conductores que quieran asegurar un vehículo según su “forma de conducir” y por ende asignar una tarifa a cada categoría.

Con respecto a la detección de anomalías, esta deberá ponerse en producción para que la empresa pueda empezar a detectar siniestros anómalos y hacerles una revisión más exhaustiva en busca de fraudes. De esta manera, se empezará a evaluar si el algoritmo de detección de anomalías en siniestros es útil o no para el problema de encontrar fraudes en reclamaciones.

Bibliografía

1. Brockett Patrick L, Xia Xiaohua, Derrig Richard A. (1998). Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance, 65(2), 245-274.

2. Chanfreut, P., Maestre, J. M., & Camacho, E. F. (2021). A survey on clustering methods for distributed and networked control systems. Annual Reviews in Control, 52, 75-90. https://doi.org/10.1016/j.arcontrol.2021.08.002

3. Espriella, C. de la. (2012). Fraude en seguros: Una aproximación al caso colombiano. Recuperado de [URL: https://www.fasecolda.com/cms/wp-content/uploads/2021/08/Fraude-en-seguros.pdf]

4. Fasecolda. (2023). Estadísticas por ramo. Recuperado de https://www.fasecolda.com/fasecolda/estadisticas-del-sector/estadisticas-por-ramo/

5. Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1), 58-75. https://doi.org/10.1016/j.jfds.2016.03.001

6. Reese, C.A., & Pash-Brimmer, A. (2009, julio). North Central Texas pay-as-you-drive insurance pilot program. En Proceedings of the Transportation, Land Use, Planning and Air Quality Conference, Denver.

7. Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2016). Innovative insurance schemes: pay as/how you drive. Disponible en [URL: http://creativecommons.org/licenses/by-nc-nd/4.0/]

8. Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. Annals of Data Science, 2(2), 165-193. https://doi.org/10.1007/s40745-015-0040-1