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 de 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 clásico de estimación de tarifas de seguros de automóviles bajo panoramas de riesgos no documentados, por lo cual el nuevo modelo depende principalmente de las características de los vehículos y sus propietarios lo que propende por 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, agrupación como k-means y 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.

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.

Materiales y Métodos

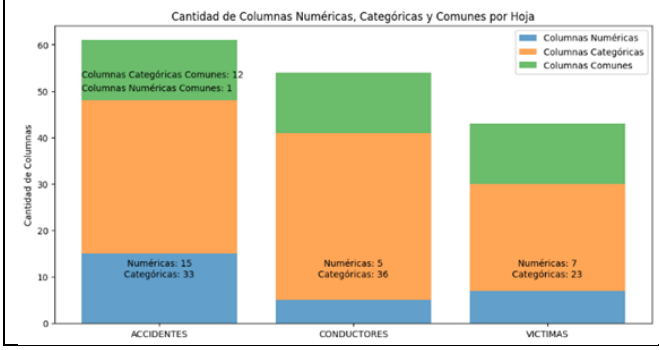
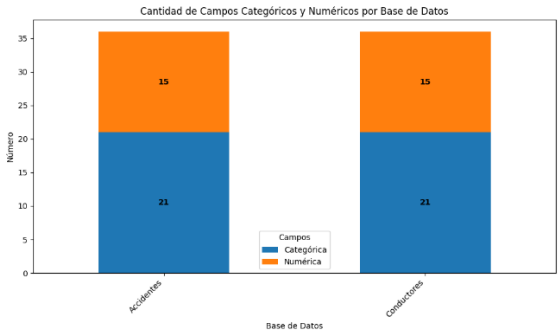
Con el fin de resolver el problema planteado se emplean los datos de https://www.movilidadbogota.gov.co/web/simur correspondiente a la consolidación de siniestros viales ocurridos durante el año 2019 en la ciudad de Bogotá, incluyendo información detallada de cada evento, ubicación geográfica, vehículos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

víctimas. El dataset original incluye tres tablas para Accidentes (34990, 50), Conductores (66179, 42) y Victimas (9465, 31), en cada uno de ellos se incluyeron variables categóricas y numéricas.

Para abordar este proyecto, se realizó el preprocesamiento de los datos; teniendo presente que se contaban con tres tablas iniciales (Accidentes, Víctimas y Conductores) se realizaron las imputaciones de datos a los registros vacíos, aplicación de sumatorias y otros operaciones algebraicas obteniendo así dos tablas transformadas (Accidentes y Conductores) se realizó entonces la identificación de las columnas de interés en los datasets y la construcción de las ETL de las mismas; se dejaron columnas categóricas y numéricas de los Conductores y aquellas que podrían tener asociación con la gravedad del accidente; lo que permite alinear al objetivo propuesto de segmentación de conductores de acuerdo con su siniestralidad.

Adicionalmente se realiza la construcción de la ETL de accidente; imputando los valores faltantes y seleccionando las columnas que puedan permite identificar posibles fraudes con base en siniestros anómalos; agregando los datos calculados de víctimas y conductores implicados en estos.

A continuación, podemos evidenciar los resultados de los procesamientos realizados en los datasets y los ETL finales con sus características principales:

Datos iniciales:	**ETL finales:**
	

Las estadísticas de los datasets finales se encuentran descritos dentro de los notebooks que acompañan el presente proyecto; a manera general se trabajan datos numéricos en rangos y desviaciones altas que representan las características de los conductores y vehículos, una columna ordinal de gravedad y categóricas sobre las condiciones del momento del accidente.

Se consideran dos procesos esenciales para la preparación de los datos con el objetivo entrenar los modelos: la binarización de variables categóricas y la estandarización de las variables numéricas; pasos necesarios para garantizar la homogeneidad y la interpretabilidad de los datos, estableciendo así modelos analíticos robustos en la identificación de patrones de los conductores en situaciones de accidentes.

En este proyecto, se han empleado una variedad de técnicas de agrupación, como K-Means, Jerárquico, K-Medioides y DBSCAN, para abordar la resolución del problema planteado. Estos algoritmos permiten facilitar la identificación de patrones y características comunes en los accidentes en grupos o clústers basados en similitudes inherentes entre ellos, y en el caso de DBScan, también podemos aprovecharlo para obtener los accidentes atípicos que puedan representar posibles fraudes en reclamaciones a la compañía de seguros.

Resultados y Discusión

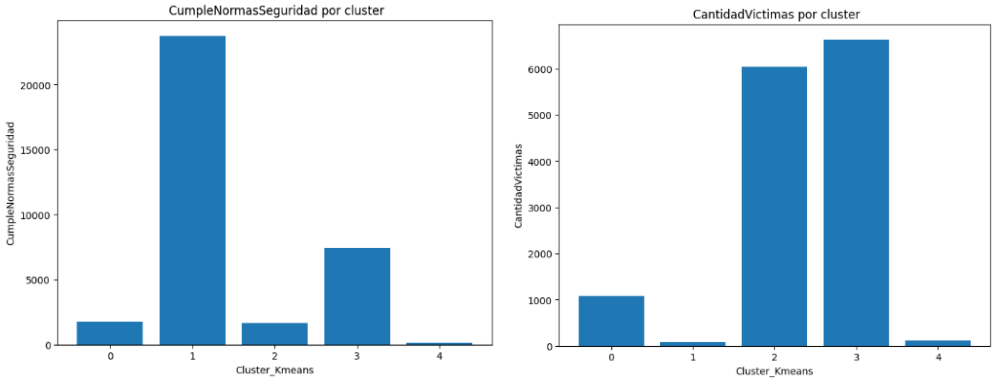
El uso de K-Means o K-Medias permite la agrupación de accidentes en clústeres en función de sus características, la gravedad del accidente, las víctimas, el tipo de vehículo involucrado y las condiciones

de cumplimiento de algunas normas de tránsito. Los algoritmos K-Medioides y Jerárquico no se pudieron correr ya que el costo computacional requerido excedía nuestra capacidad inclusive cuando utilizamos el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos y servidores en nube con buena capacidad como Google Colab.

Debido a lo anterior, K-Means se convirtió en el algoritmo principal en nuestro análisis, y aunque inicialmente se realizaron esfuerzos exhaustivos para encontrar los parámetros óptimos mediante enfoques matemáticos como prueba de codo y silueta, los resultados obtenidos no fueron concluyentes. Los métodos basados en matemáticas y métricas de evaluación no proporcionaron una solución definitiva para la selección de parámetros, ya que la naturaleza de los datos de accidentes_conductores presentaba ciertas complejidades que dificultaban la determinación precisa de los mejores parámetros.

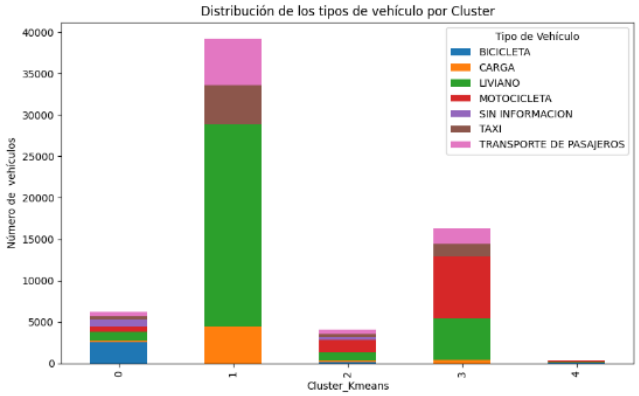
Tras desarrollar el algoritmo se logró identificar ***cinco clústeres*** que resaltan la gravedad de los accidentes al considerar múltiples variables, que abarcan las características del accidente, como la embriaguez, la presencia de peatones y la cantidad de víctimas. Este enfoque nos proporciona una comprensión concisa de la diversidad en la gravedad de los accidentes y con base en la asignación de clústers podríamos definir una tarifa según las consecuencias de cada uno, posteriormente se cruza estos clústers con la información y características originales de los conductores para identificar las posibles agrupaciones que deberían tener una u otra tarifa.

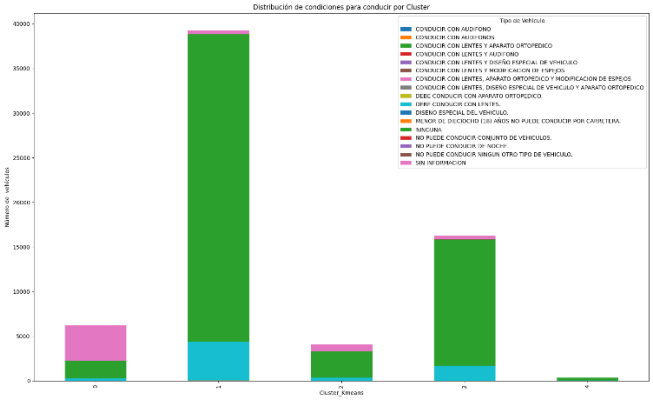
Al analizar los clústeres en función de la cantidad de víctimas y aquellos accidentes en los que se cumplen las normas de seguridad, se refleja como 2 de los 5 clústeres (clústeres 2 y 3) representan un mayor riesgo al contar con una mayor cantidad de víctimas, dentro de las que se incluyen heridos, casos de hospitalización y muertos. En contra parte, se identifica un clúster que representa menores riesgos, al estar asociado con los accidentes en los que más se cumplen las normas de seguridad que a su vez son aquellos en los que menos víctimas se presentan.



Dados estos resultados, para el negocio es relevante entender las características de los conductores que conllevan a que se registren mayor cantidad de víctimas y entender las similitudes que se encuentran en los clústeres 2 y 3 y diferencias frente al clúster 1. Lo anterior, facilitará la generación y asignación de tarifas a cada tipo de conductor.

Tras realizar el merge de los clústeres identificados con el algoritmo con la base de conductores, se encontró que el tipo de vehículo mediante el cual se movilizan los conductores es determinante a la hora de evaluar el riesgo de accidentalidad al cual están expuestos, de forma que aquellos que se movilicen por medio de motocicletas, son más propensos a no cumplir normas de seguridad y a tener accidentes en los que se presenten víctimas, ya sean heridos o muertos. Esto lo podemos observar en el gráfico de distribución de los tipos de vehículo por clúster, en el cual las motocicletas tienen una alta participación en el clúster 3 y 2, sin tener presencia en el número 1.

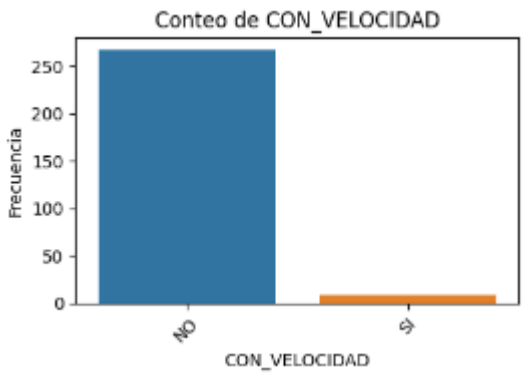


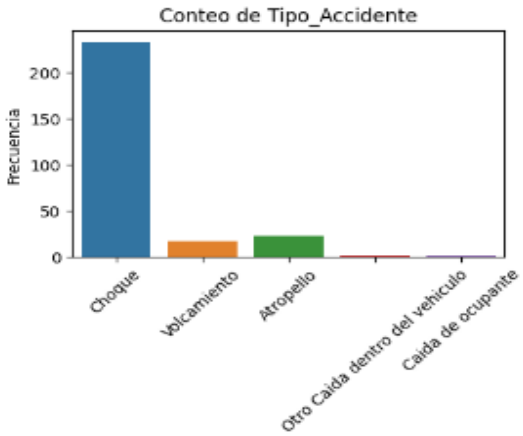


Por otro lado, al analizar los resultados de los clústeres contra las condiciones que tienen algunos conductores a la hora de conducir, tales como si deben usar lentes, conducir con audífonos, o no conducir de noche, se evidencia que estas no son representativas a la hora de determinar el riesgo que puede presentar un conductor ante accidentes, de forma que, por ejemplo, aquellos que usan lentes se encuentran tanto en el clúster más seguro (1) como en el de más víctimas (3).

Por último, DBSCAN lo empleamos para la detección de outliers y la identificación de accidentes singulares, lo que contribuye al cumplimiento del segundo objetivo del proyecto: la identificación de fraudes en las reclamaciones de accidentes de tránsito. Durante la implementación del DBScan tuvimos que iterar bastante en cuanto a los N vecinos para obtener un “eps” lo suficientemente grande con el knee locator que no nos dejará demasiados outliers y de igual manera iteramos varios min_samples con el mismo objetivo. Esto se hizo con el fin de encontrar siniestros atípicos que fueran menos del 1% de la totalidad de datos para que fuese práctico para la compañía evaluar esta cantidad de siniestros con más detalle e identificar si realmente son fraude o no. Finalmente se encontraron 276 registros atípicos que corresponden aproximadamente al 0.8% del total de los datos.

Para medir la efectividad de los posibles fraudes se realizó un análisis descriptivo de negocio que permitió 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).





Al analizar estas características encontramos que la mayoría de los siniestros atípicos si tienen características comunes y fácilmente simulables para realizar una reclamación fraudulenta al seguro, por ejemplo, encontramos que la mayoría de estos siniestros atípicos no fueron ocasionados por altas velocidades, fueron choques de latas (solo daños) o volcamientos, la colisión fue contra otros vehículos

u objetos fijos y normalmente no hubo heridos. Esto tiene mucho sentido pues para realizar una reclamación cualquiera podría simular un siniestro con estas características sin mayores riesgos.

Sin embargo, para validar la calidad y utilidad real de 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 realmente útil o no para el problema de encontrar fraudes en reclamaciones.

Conclusión

En conclusión, este estudio ha enfrentado una serie de desafíos computacionales que han influido en la calidad de los resultados obtenidos. Las dificultades computacionales surgieron debido a la complejidad de los algoritmos utilizados, como el K-Mediodes y el clústering Jerárquico, que, a pesar de realizar un ajuste minucioso de parámetros para alcanzar un nivel aceptable de rendimiento, nos fue imposible lograr que el algoritmo corriera en nuestras máquinas o Google Colab. Además, la inconsistencia de la información suministrada por el distrito en relación con la gravedad de los accidentes y otros detalles ha planteado obstáculos para la precisión de nuestro análisis.

La falta de datos también ha sido una limitación significativa en este estudio, ya que la ausencia de información detallada sobre ciertas variables relevantes ha dificultado nuestra capacidad para identificar posibles características y patrones con confianza, como grado de alcoholemia, días de incapacidad u otros que podrían haber apoyado el presente proyecto. Además, hemos observado un sesgo en los datos, lo que significa que ciertos grupos o tipos de accidentes pueden estar subrepresentados o sobrerrepresentados en nuestra muestra, lo que potencialmente distorsiona nuestras conclusiones; adicionalmente teniendo en cuenta que la muestra de vehículos circulantes está desbalanceada puesto que según el Registro Único Nacional de Tránsito (Runt) [2023], en 2019 circulaban en Bogotá un total de 15,3 millones de vehículos, de los cuales 7 eran motos, 4,5 eran carros, 2,8 millones eran camionetas y 1 millones; por lo que los datos están centrados en aquellos que han sufrido un accidente.

Pese a las dificultades descritas anteriormente, por medio del algoritmo K-means, se logró obtener resultados que permiten identificar características que identifican a los conductores que están asociados a mayores riesgos en las vías, y así mismo descartar características que permitan categorías tarifarias en los seguros. En línea con esto, aquellos conductores que se encuentren transitando por medio de motocicletas son los que mayor riesgo representaran al negocio a la hora de hacer efectivo un seguro.

Con respecto a los resultados obtenidos en la identificación de anomalías con DBScan concluimos que los registros de accidentes marcados como anómalos si tienen características que indican una alta posibilidad de que sean fraudes, sin embargo, como se mencionó en la sección de resultados es necesario poner el algoritmo en producción de tal forma que la empresa revise e investigue en mayor detalle los siniestros que arroje el algoritmo y pueda realmente determinar si son fraudes o no, y por ende si el algoritmo genera valor o no a la organización con el objetivo de detectar y prevenir fraudes en reclamaciones de seguros.

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. Registro Único Nacional de Tránsito (Runt). (2019). Parque automotor de Bogotá en 2019. Recuperado de https://runt.com.co/Estadisticas/Consultas/Parque-Automotor/

8. 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/]

9. 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