Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Estudios de Postgrado

ANÁLISIS DE PATRONES DE LESIONADOS Y FALLECIDOS EN ACCIDENTES DE TRÁNSITO EN GUATEMALA (2015-2023) MEDIANTE MINERÍA DE DATOS

Brandon René Portillo González

Carnet: 999011994

Repositorio: https://github.com/usac201612398/DataMining

Guatemala, noviembre 2025

INDICE DE CONTENIDO

1. RESUMEN 1
2. INTRODUCCION 2
3. METODOLOGIA DE LA INVESTIGACION 3
   3.1. Datos y Preparación 3
   3.2. Técnicas de análisis 3
4. PRESENTACION DE RESULTADOS 4
   4.1. Reglas de asociación 4
      4.1.1. Apriori aplicado a nivel nacional 4
      4.1.2. Apriori aislando incidencias en la ciudad capital 6
      4.1.3. FP-Growth (Fallecidos y lesionados) 8
      4.1.4. FP-Growth (Fallecidos) 9
   4.2. Análisis de clúster (K-Means) 10
      4.2.1. El método del codo 10
      4.2.2. Variables con mejor separación euclidiana 11
      4.2.3. Parametrización de clusters 12
5. DISCUSION DE RESULTADOS 15
   5.1. Limitaciones de la investigación 15
6. CONCLUSIONES 16
   6.1. Segmentación por edad de la población afectada 16
   6.2. Motocicleta como factor crítico 16
   6.3. Patrones geográficos y temporales 16
   6.4. Tipo de evento más peligroso 17
7. RECOMENDACIONES 18
8. REFERENCIAS 19

INDICE DE TABLAS

Tabla 1. Reglas de asociación obtenidas mediante Apriori para análisis nacional 4
Tabla 2. Codificación de depto_ocu y su descripción popular 5
Tabla 3. Reglas de asociación obtenidas mediante Apriori para casos centrados en la ciudad capital..... 7
Tabla 4. Reglas de asociación obtenidas mediante FP-Growth para casos centrados en la ciudad capital..... 8
Tabla 5. Reglas de asociación obtenidas mediante FP-Growth para muertes registradas en la ciudad capital..... 9
Tabla 6. Distribución de medias para centroides según pares de variables 11

INDICE DE FIGURAS

Figura 1. Visualización de reglas de asociación con mayor densidad de incidencias. 6
Figura 2. Determinación de clusters mediante el método del codo. 10
Figura 3. Fallecidos/Lesionados vs Edad..... 12
Figura 4. Tipo-vehículo vs Edad..... 13
Figura 5. Mes vs Tipo-vehículo 13
Figura 6. Tipo-vehículo vs Tipo-evento 14

1. RESUMEN

En la primera fase de este proyecto se analizó datos históricos de accidentes de tránsito reportados por la Policía Nacional Civil (PNC) de Guatemala entre 2015 y 2023, con énfasis en lesionados y fallecidos. Se aplicaron técnicas de minería de datos como Apriori, FPGrowth y K-means para identificar patrones y relaciones entre variables como edad, sexo, tipo de vehículo, tipo de evento y departamento de ocurrencia.

Los resultados mostraron que los jóvenes entre 0 y 33 años son los más afectados, especialmente en motocicletas, y que los eventos de colisión, choque y vuelco son los más frecuentes. El departamento de Guatemala concentra la mayor incidencia. Se concluye que es necesario implementar políticas de prevención dirigidas a grupos de riesgo y vehículos específicos.

2. INTRODUCCION

Los accidentes de tránsito son una causa significativa de lesionados y fallecidos en Guatemala. Este estudio utiliza minería de datos para explorar patrones ocultos en los registros de la PNC, con el objetivo de identificar factores de riesgo y contribuir a la toma de decisiones en seguridad vial. Se aplicaron algoritmos de asociación y clustering para analizar datos de 2015 a 2023, excluyendo años con información incompleta o inconsistente (no disponían el formato de manera uniforme).

3. METODOLOGIA DE LA INVESTIGACION

3.1. Datos y Preparación

Se consolidaron datos de 8 archivos Excel (2015-2020, 2022-2023) y 1 archivo SPSS (2021), totalizando más de 75,000 registros. Las variables seleccionadas incluyeron: año, mes, día, departamento, sexo, edad, condición (lesionado/fallecido), tipo de vehículo y tipo de evento. Se eliminaron registros con valores desconocidos (ej.: edad = 999, sexo = 9, tipov_veh=99) para reducir ruido.

3.2. Técnicas de análisis

- **Apriori y FPGrowth:**

  Para generar reglas de asociación con soporte ≥ 0.2 y confianza ≥ 0.5.

- **K-means:**

  Para clustering basado en variables numéricas, con normalización previa y determinación de 5 clusters mediante el método del codo.

- **Visualización:**

  Gráficos de dispersión, barras y redes para interpretar resultados.

4. PRESENTACION DE RESULTADOS

4.1. Reglas de asociación

4.1.1. Apriori aplicado a nivel nacional

Se realiza el primer análisis para ello se discrimina sobre 10 variables con el propósito de evitar el sobre procesamiento y disminuir el ruido debido a la sobrecarga de categorías dentro del set de datos. Por ejemplo, la marca de los vehículos y los municipios de incidencia.

Se presentan las 4 reglas con mayor valor en “Confidence”, “Suppor” y “Lift”, las cuales nos indican la relevancia de los patrones que encontró el modelo Apriori con la condición de que mientras mayor sean estos parámetros más alarmantes pueden ser las observaciones dependiendo de si se espera o no un evento determinado.

Tabla 1.

Reglas de asociación obtenidas mediante Apriori para análisis nacional

Lhs	Rhs	Support	Confidence	Coverage	Lift	Count
[1] {edad_per=[23,34), fall_les=[2,3]}	=> {tipo_eve=[1,4)}	0.2004942	0.7468248	0.2684622	1.12075	16876
[9] {edad_per=[0,23), tipo_eve=[1,4)}	=> {fall_les=[2,3]}	0.2004823	0.8923378	0.2246709	1.08176	16875
[21] {depto_ocu=[1,11), tipo_eve=[1,4)}	=> {fall_les=[2,3]}	0.3801977	0.876263	0.4338854	1.06227	32002
[37] {fall_les=[2,3], tipo_veh=[4,99]}	=> {tipo_eve=[1,4)}	0.4053248	0.7049404	0.5749774	1.05789	34117

Nota. Elaborado en Word 365 y obtenido de procesamiento mediante el lenguaje R versión 4.5.1.

A continuación de desglosa el análisis correspondiente para cada una de los patrones observados:

1) Entre 23 y 33 años, ya sea fallecido o lesionado, el tipo de evento de reporte de accidente tiende a ser colisión, choque o vuelco con alta probabilidad (74 % de las veces) y una frecuencia del 20%.
2) Entre las personas jóvenes reportadas existe una mayor probabilidad de ocurrencia de lesiones o muertes debido a colisiones, choques o vuelcos (89% de las veces) y una frecuencia del 22%
3) En ciertos departamentos que probablemente estén entre los más poblados (1 al 10), el 87 % de los casos terminan con lesionados/fallecidos, debido a choques, colisiones o vuelcos. Esa es una regla muy significativa porque tiene soporte alto (38 %) y buena confianza.

Tabla 2.

Codificación de depto_ocu y su descripción popular

Código	Departamento
1	Guatemala
2	El progreso
3	Sacatepéquez
4	Chimaltenango
5	Escuintla
6	Santa Rosa
7	Sololá
8	Totonicapán
9	Quetzaltenango
10	Suchitepéquez

Nota. Diccionario para la variable depto_ocu, elaborado en Word 365.

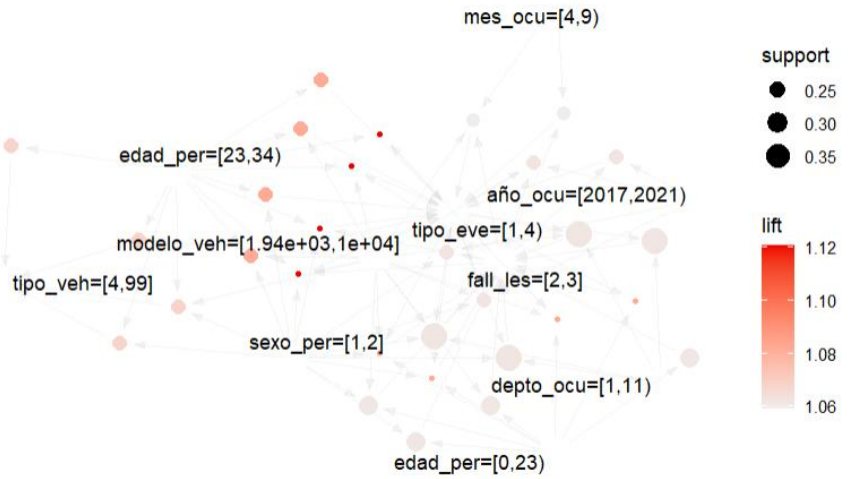
4) Esta regla nos indica que estos accedentes de tránsito donde hay lesionados/fallecidos utilizan vehículos que van desde las motocicletas o desconocidos con una probabilidad del 70% y una frecuencia del 40%.

Con estas reglas generales podemos hacer discriminación por género, grupos de edad, área geográfica entre otros parámetros que permitan indagar más en la información recolectada de los años 2017 al 2021.

La siguiente imagen representa gráficamente la distribución de cada uno de los focos donde existe mayor incidencia, a través de la librería arulesViz

Figura 1.

Visualización de reglas de asociación con mayor densidad de incidencias.



Nota. Elaboración propia. Obtenido mediante biblioteca arulesViz procesando mediante el lenguaje R versión 4.5.1.

4.1.2. Apriori aislando incidencias en la ciudad capital

Fue necesario segmentar el departamento de Guatemala para ver los patrones en esta área metropolitana ya que es el lugar donde hay más tránsito a nivel nacional, además se excluyó no solo la marca sino que también el modelo del vehículo con el que se reportaron los hechos debido a que inyectaba demasiado ruido en las reglas y se hizo pruebas reduciéndolas de 200 a 94 en las cuales se describe lo siguiente:

Tabla 3.

Reglas de asociación obtenidas mediante Apriori para casos centrados en la ciudad capital

Lhs	Rhs	Support	Confidence	Coverage	Lift	Count
[5] {edad_per=[0,23)}	=> {tipo_veh=[4,99]}	0.21894	0.75793	0.28886	1.024	6287
[92] {edad_per=[23,33), tipo_veh=[4,99], tipo_eve=[1,5)}	=>{sexo_per=[1,2]}	0.20898	1	0.20898	1	6001
[79] {edad_per=[23,33), tipo_eve=[1,5)}	=>{tipo_veh=[4,99]}	0.20898	0.81106	0.25766	1.096	6001
[20] {año_ocu=[2018,2021)}	=>{tipo_eve=[1,5)}	0.22249	0.66476	0.33469	1.029	6389

Nota. Elaborado en Word 365 y obtenido de procesamiento mediante el lenguaje R versión 4.5.1.

1. Según esta regla, para el caso de los accidentes reportados en el departamento de Guatemala, existe una tendencia de que los jóvenes que conducen motocicletas sean los perjudicados con una probabilidad del 75% y en un 21% de las veces.
2. Se tiene una probabilidad del 100% que los fallecimientos/lesionados sean reportados por conducir en moto u otros vehículos cuando el evento es una colisión, independientemente que sea hombre o mujer cumpliéndose el 20% de las veces.
3. Dado que la persona tiene entre 23 y 33 de edad, y se accidento mediante choque o colisión hay probabilidad del 81% que se cumpla el 20% de las veces haya sucedido en una motocicleta o vehículo desconocido
4. Del 2018 al 2020 hay una probabilidad del 66% que los accidentes ocurrieran a través de colisiones choques o volteos con una frecuencia del 22% de las veces.

4.1.3. FP-Growth (Fallecidos y lesionados)

Tomando en cuenta a la población del departamento de Guatemala y considerando tanto hombres como mujeres, asimismo tanto lesionados como fallecidos se obtiene las siguientes reglas de asociación:

Tabla 4.

Reglas de asociación obtenidas mediante FP-Growth para casos centrados en la ciudad capital

rules	support	confidence	lift	count
{sexo_per=[1,2],tipo_veh=[1,4)} => {fall_les=[2,3]}	0.2299763	0.88478	1.00895	6604
{sexo_per=[1,2],edad_per=[0,23)} => {fall_les=[2,3]}	0.2651483	0.9179	1.04672	7614
{mes_ocu=[1,5),tipo_veh=[4,99]} => {fall_les=[2,3]}	0.2118331	0.875	0.9978	6083
{día_ocu=[11,21),tipo_veh=[4,99]} => {fall_les=[2,3]}	0.2178228	0.87299	0.99551	6255

Nota. Elaborado en Word 365 y obtenido de procesamiento mediante el lenguaje R versión 4.5.1.

Existe una fuerte correlación para estas 4 reglas obtenidas mediante FP-Growth de que los accidentes ya sea que repercutan en muertes o lesiones sean aquellos casos en los que las personas se conducen en motocicleta o vehículos desconocidos, independientemente de que si son hombres o mujeres, cumpliéndose el 25% de las veces con altas probabilidades.

En los primeros cinco meses del año se estima que un accidente suceda en moto en la metrópoli con un 87.5% de probabilidad y además también existe una regla que correlaciona que esa misma probabilidad e indica que los incidentes pueden suceder a mediados de mes.

El dato alarmante es que las incidencias suceden con mayor probabilidad para hombres o mujeres entre los 0 y 23 años, es decir hay una población joven significativa.

4.1.4. FP-Growth (Fallecidos)

Se procede a analizar los casos que concluyen en muertes debido a accidentes de tránsito focalizados en el departamento de Guatemala, con lo cual no solo se redujo la cantidad de reglas a 100, sino que se puede conocer las estadísticas que respaldan el peor de los casos, de las cuales se destacan las siguientes 4:

Tabla 5.

Reglas de asociación obtenidas mediante FP-Growth para muertes registradas en la ciudad capital

rules	support	confidence	lift	count
{sexo_per=[1,2],edad_per=[23,33),tipo_eve=[1,5)} => {tipo_veh=[4,99]}	0.2122319	0.81031	1.09833	5344
{tipo_veh=[4,99]} => {tipo_eve=[1,5)}	0.5151708	0.69828	1.0484	12972
{mes_ocu=[9,12],sexo_per=[1,2]} => {tipo_veh=[4,99]}	0.2589754	0.73742	0.99953	6521
{día_ocu=[21,31],sexo_per=[1,2]} => {tipo_veh=[4,99]}	0.246942	0.73438	0.99541	6218

Nota. Elaborado en Word 365 y obtenido de procesamiento mediante el lenguaje R versión 4.5.1.

1) Para el caso de muertes en el departamento de Guatemala se aprecia que para personas entre los 23 y 32 años tienen una probabilidad de 81% de perder la vida en un accidente de tipo colisión, choque, vuelco o caída ya sea en motocicleta o medio de transporte desconocido.

2) Existe una probabilidad del 69% cumpliéndose el 51% de los casos que las personas que pierdan la vida en incidentes de tránsito de tipo colisión, choque, caída, vuelco sea quien conduzca una motocicleta.

3) Otra de las reglas que puede valorarse es que en los meses de septiembre a diciembre van a existir muertes debido a accidentes por colisionar, caer, vuelco o choque si se conduce en motocicleta o algún otro vehículo desconocido con una probabilidad del 73% y cumpliéndose el 25% de las veces.

4) Otra regla importante es cuando se acerca el fin de mes, independiente si es hombre o mujer en Guatemala suceden muertes con una

probabilidad del 73% cumpliéndose el 25% de las veces si se conduce mediante motocicleta.

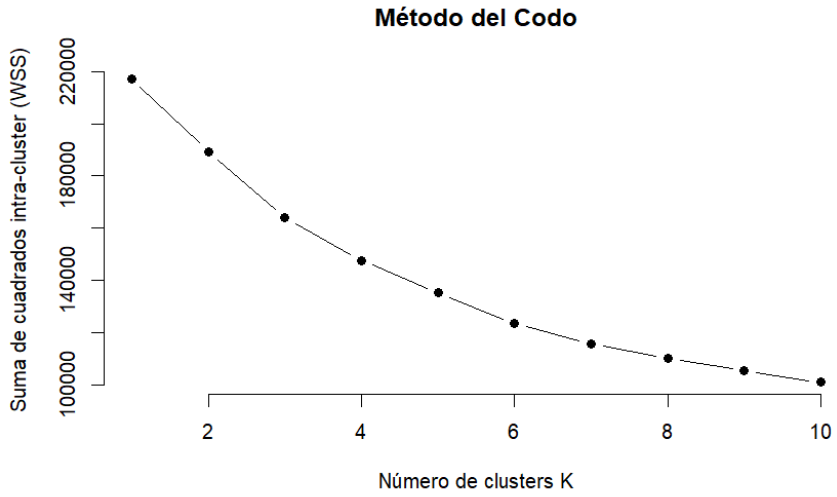
4.2. Análisis de clúster (K-Means)

4.2.1. El método del codo

Se utiliza el dataset que contiene alrededor de ocho variables, las cuales nos permitieron indagar en la búsqueda de patrones dentro del set de datos, con la finalidad de determinar las correlaciones entre las variables (previamente escaladas). Adicional al contenido visto en clase se realizó el análisis del codo para determinar cuantos centroides encontramos optimando dentro de nuestro set de datos.

Figura 2.

Determinación de clusters mediante el método del codo.



Nota. Elaboración propia. Obtenido mediante biblioteca ggplot2 procesando mediante el lenguaje R versión 4.5.1.

4.2.2. Variables con mejor separación euclidiana

Existen distintos pares de variables que muestran cada separación entre los clusters que se generaron con el modelo K-means. Para cada posible combinación de dos variables, se mide qué tan separados están los centroides de los clusters en ese plano de lo que se obtiene que cuanto mayor sea la distancia promedio entre los centroides, más visualmente distinguibles serán los clusters en ese par de variables.

Las primeras filas son las mejores combinaciones para graficar, porque los centroides de los clusters están más separados, por lo tanto, es más probable que los grupos se vean bien en el gráfico.

Tabla 6.

Distribución de medias para centroides según pares de variables

var1 <chr>	var2 <chr>	mean_centroid_distance <dbl>
edad_per	tipo_veh	5.3332452
tipo_veh	tipo_eve	4.7835745
fall_les	tipo_veh	3.8777279
sexo_per	tipo_veh	3.8776358
año_ocu	tipo_veh	3.7544841
mes_ocu	tipo_veh	3.7283640
día_ocu	tipo_veh	3.6772571
edad_per	tipo_eve	3.3956039
año_ocu	edad_per	2.8199187
sexo_per	edad_per	2.7154184
mes_ocu	edad_per	2.6324846
día_ocu	edad_per	2.6283058
año_ocu	tipo_eve	1.9842788
fall_les	tipo_eve	1.9002564
sexo_per	tipo_eve	1.8949802
mes_ocu	tipo_eve	1.7751661
día_ocu	tipo_eve	1.7382572
año_ocu	fall_les	0.8930088
año_ocu	sexo_per	0.8859349
sexo_per	fall_les	0.7894155
año_ocu	mes_ocu	0.6732561
año_ocu	día_ocu	0.6163823
mes_ocu	sexo_per	0.5774084
mes_ocu	fall_les	0.5483737
día_ocu	sexo_per	0.5244649
día_ocu	fall_les	0.4657196
mes_ocu	día_ocu	0.2454370

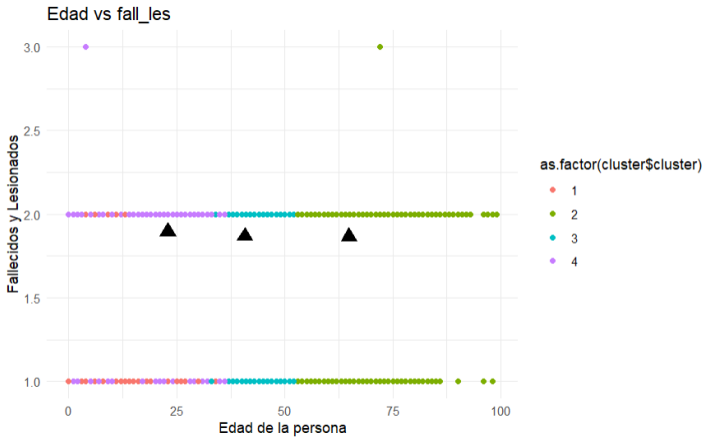
Nota. Elaborado en Word 365 y obtenido de procesamiento mediante el lenguaje R versión 4.5.1.

4.2.3. Parametrización de clusters

Se realiza un análisis de los cuatro clusters de acuerdo con el método del codo y las correlaciones que mejor existen entre cada una de las variables.

Figura 3.

Fallecidos/Lesionados vs Edad

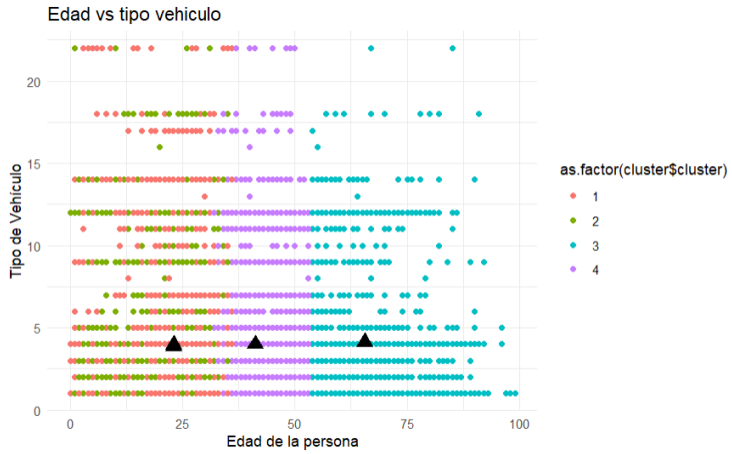


Nota. Elaboración propia. Obtenido mediante biblioteca ggplot2 procesando mediante el lenguaje R versión 4.5.1.

Según el patrón que exhibe la Figura 3, hay mas concentración tanto de heridos más las personas menores de 34 años y mayores de 50

Figura 4.

Tipo-vehículo vs Edad

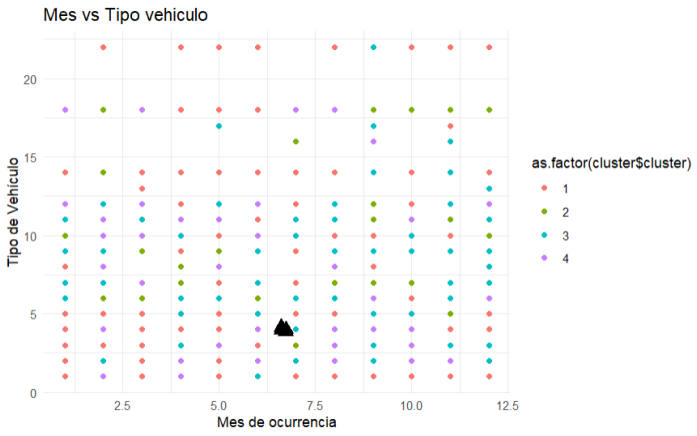


Nota. Elaboración propia. Obtenido mediante biblioteca ggplot2 procesando mediante el lenguaje R versión 4.5.1.

Con base en lo que describe la Figura 4, se concluye que independientemente de la edad hay más tendencia de que los accidentes sucedan en vehículos comunes tales como automóviles, camionetas, pick ups, motocicletas y camiones.

Figura 5.

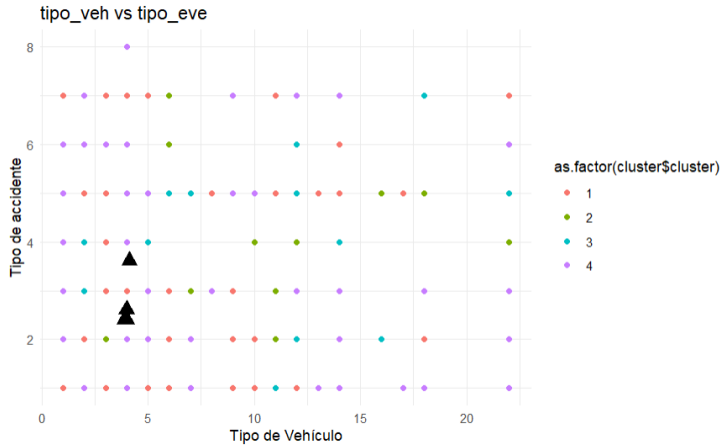
Mes vs Tipo-vehículo



Nota. Elaboración propia. Obtenido mediante biblioteca ggplot2 procesando mediante el lenguaje R versión 4.5.1.

Figura 6.

Tipo-vehículo vs Tipo-evento



Nota. Elaboración propia. Obtenido mediante biblioteca ggplot2 procesando mediante el lenguaje R versión 4.5.1.

En general las visualizaciones revelaron que:

- La edad y el tipo de evento están correlacionados, con jóvenes involucrados en colisiones.
- Motocicletas están sobrerrepresentadas en eventos con lesionados/fallecidos.
- Los meses de septiembre a diciembre y fines de mes concentran incidentes graves.

5. DISCUSION DE RESULTADOS

Los resultados indican que los jóvenes, especialmente aquellos entre 0 y 33 años, son los más vulnerables en accidentes de tránsito, con motocicletas como vehículos de alto riesgo. La alta incidencia en el departamento de Guatemala sugiere la necesidad de campañas de prevención en áreas urbanas.

Las reglas de asociación y clustering revelaron que los tipos de evento "colisión", "choque" y "vuelco" son los más peligrosos, coincidiendo con hallazgos previos (OMS, 2018). La estacionalidad (meses y días específicos) podría relacionarse con factores climáticos y comportamientos sociales.

5.1. Limitaciones de la investigación

- Falta de datos para 2024-2025 e inconsistencia en formatos previos a 2015.
- Exclusión de datos de Provial por estructura dispar.
- Posible subregistro en variables como tipo de vehículo y modelo.

6. CONCLUSIONES

La minería de datos permitió identificar patrones críticos en accidentes de tránsito guatemaltecos.

6.1. Segmentación por edad de la población afectada

- Los jóvenes (15-30 años) constituyen el grupo más vulnerable, el cual está asociado con motocicletas y eventos de colisión.
- Los adultos jóvenes (23-34 años) muestran la mayor probabilidad de participar en accidentes graves.
- Cada grupo etario presenta patrones distintivos en tipo de vehículo y evento

6.2. Motocicleta como factor crítico

- Las motocicletas representan el mayor riesgo, con probabilidades del 70-80% en accidentes con lesionados/fallecidos.
- Las motocicletas representan el mayor riesgo, con probabilidades del 70-80% en accidentes con lesionados/fallecidos

6.3. Patrones geográficos y temporales

- Hay patrones estacionales, con mayor incidencia en meses específicos y fines de mes
- Los primeros cinco meses del año muestran probabilidades del 87.5% para accidentes en motocicletas

6.4. Tipo de evento más peligroso

- Colisiones, choques y vuelcos representan los eventos con mayor probabilidad de resultar en lesionados o fallecidos
- Existe una correlación del 89% entre jóvenes y estos tipos de eventos graves

7. RECOMENDACIONES

- Fortalecer regulaciones para motocicletas y conductores jóvenes entre 15 y 30 años.
- Implementar campañas de concientización en departamentos prioritarios y en conductores de motocicletas.
- Utilizar estos hallazgos para optimizar recursos de respuesta a emergencias.

8. REFERENCIAS

Organización Mundial de la Salud (OMS). (2018). Informe sobre la situación mundial de la seguridad vial.

Wickham, H., & Grolemund, G. (2016). R for Data Science. O'Reilly Media.

Guatemala, I. N. (5 de Noviembre de 2025). *INE*. Obtenido de INE: https://www.ine.gob.gt/bases-de-datos/accidentes-de-transito/

González, B. R. (7 de 11 de 2025). *GitHub*. Obtenido de usac201612398/DataMining: https://github.com/usac201612398/DataMining