PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR

FACULTAD DE INGENIERIA

MAESTRÍA EN CIENCIA DE DATOS

TESIS

"Aplicación de técnicas de agrupamiento para caracterizar patrones de siniestros viales en Ecuador en el año 2021"

AUTOR: Edwin Alcides Maza Jara, Ing

DIRECTOR: Alfredo Calderón Serrano, Ing., MSc.

Quito - 2023

1 Agradecimientos

Deseo agradecer a todas las personas que me han brindado su apoyo incondicional para cumplir todos mis objetivos personales y académicos. En especial quiero agradecer a mi familia, base fundamental, que con su sacrificio personal, económico y emocional me han permitido llegar hasta el final.

Por último, agradecer a mis padres quienes continúan inspirándome a cumplir sueños y metas. A quienes debo todos los frutos cosechados.

2 Dedicatoria

Dedico este trabajo a mi esposa Andrea, por su paciencia, compresión y sobre todo por su sacrificio quien me apoyo en los momentos malos y menos malos.

A mi hijo Ezequiel, fuente de motivación e inspiración que me ha permitido dar un paso más, para crecer como persona y que este logro sirva de herramienta para guiar cada uno de sus pasos.

3 Resumen

El presente trabajo de investigación aplicada realiza un estudio sobre los patrones de siniestralidad vial del Ecuador en el año 2021, utilizando modelos no supervisados de agrupamiento de forma que se pueda caracterizar los grupos más comunes en causar accidentes viales.

Para el desarrollo de la investigación aplicada se utiliza la metodología de minería de datos CRIPS DM, la cual permite entender la problemática y establecer los objetivos a resolver en sus diferentes fases.

Finalmente se provee información adecuada de las características y grupos que más incidencia tienen en los accidentes de tránsito en Ecuador.

Palabras clave: accidente vial, modelos no supervisados, agrupamiento, clúster, CRIPS-DM

4 Abstract

This applied research work conducts a study on road accident patterns in Ecuador in the year 2021, using unsupervised clustering models to characterize the most common groups that cause road accidents.

For the development of the applied research, the CRIPS DM data mining methodology is used, which allows understanding the problem and establishing the objectives to be solved in its different phases.

Finally, adequate information is provided on the characteristics and groups that have more incidence in traffic accidents in Ecuador.

Key words: road accident, unsupervised models, clustering, clustering, CRIPS-DM.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

1. Introducción

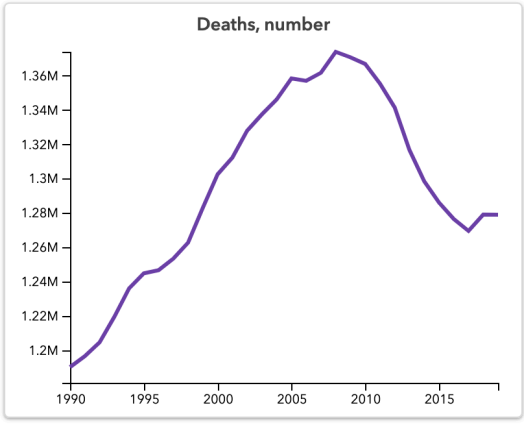
En este capítulo se describe el área de investigación aplicada que cubre el presente trabajo de titulación, Se plantea el problema a resolver, así como el contexto general de la motivación y la modelo de solución propuesto. Se formula el objetivo de la investigación aplicada, y finalmente se describe el planteamiento del trabajo.

1.1. Antecedentes

El transporte vehicular desde sus inicios ha beneficios a los seres humanos, facilitando el traslado de bienes y personas, de la misma manera el crecimiento vertiginoso del parque automotor trae consigo efectos negativos, como los sociales y ambientales. Según el proyecto de la OMS sobre la Carga Mundial de Morbilidad - CMM del 2019, los accidentes de tránsito causaron más de 1,278 millones de muertes en ese año (CMM, 2019)

Figura 1-1 Estudio de Carga Mundial de Morbilidad - CMM

Measure	Metric	Cause	Location	Age	Sex	Year	Value	Upper	Lower
Deaths	Number	Transport injuries	Global	All ages	Both sexes	2019	1,278,878.92	1,392,622.10	1,130,883.39



Nota. Número de muertes y lesiones causadas por accidentes de tránsito en el año 2019. Tomada de (Global Burden of Disease, 2019)

Según la OMS (2019), cada año los siniestros viales provocan la muerte de aproximadamente 1,3 millones de personas, siendo la 8a causa de muerte, así mismo es considerada la principal causa de mortalidad entre los niños y jóvenes de 5 a 29 años, afectando a más personas cada año que el VIH, la malaria y otras enfermedades frecuentes.

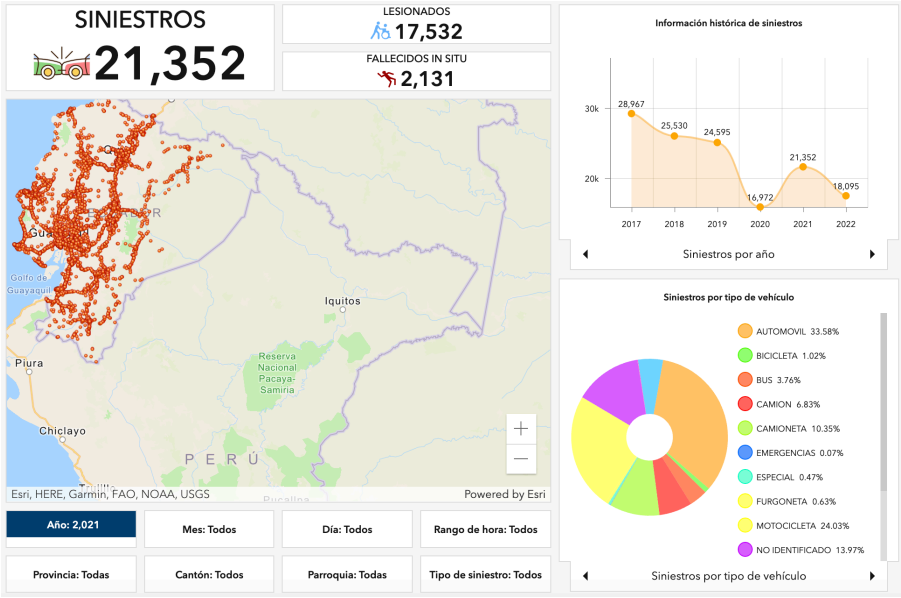
Acorde con OPS (2019), los usuarios más vulnerables en la vía son los peatones, los ciclistas y motociclistas. De la misma manera se ha informado que los siniestros viales repercuten directamente al desarrollo económico de cada país, afectando el 3% del PIB en la mayoría de los países contribuyendo como factor de pobreza. (Organización Panamericana de la Salud, Organización Mundial de la Salud, 2019).

De acuerdo la OIS (2019), los rangos de edad que concentran el 42% de fallecidos en América Latina están entre los 15 – 24 años y 25 -34 años, estos representan una tasa de mortalidad del 2.0 pada cada uno de los grupos etarios. (Observatorio Iberoamericano de Seguridad Vial, 2019)

Por otra parte, las estadísticas presentadas por la ANT (2020), muestran que el contexto vial en Ecuador es más complejo en relación con América Latina. En el año 2020, se reportaron 16.972 siniestros viales, de los cuales 13.099 fueron reportadas como lesiones graves y 1.591 reportados como fallecimientos en sitio.

Igualmente, para el año 2021 ANT del Ecuador, registró 21.352 siniestros de tránsito cuya tasa de mortalidad por cada 100.000 habitantes es de 12, lo cual representa un incremento de 2.9 puntos con respecto al año 2020 (Agencia Nacional de Transito del Ecuador, 2021).

Figura 1-2: Siniestralidad Vial del Ecuador en el año 2021



Nota. Estadísticas de siniestralidad vial del ecuador en el año 2021. Tomado de (Agencia Nacional de Tránsito, 2021)

Así mismo la ONU (2020), promueve la Segunda Década del Plan Mundial Para el Decenio De Acción Para la Seguridad Vial 2021-2030, cuyo principal objetivo es reducir al 50% de muertes y lesiones caudas por siniestros viales para el 2030. Ecuador en consonancia con el objetivo de la OMS, y alineado con los pilares para la seguridad vial, elabora varios planes de acción como: mejoramiento de la infraestructura vial, gestión de la seguridad vial, movilidad sostenible y desarrollo urbano, seguridad de los usuarios vulnerables. (Organización de Naciones Unidad, 2020)

Las entidades públicas del Ecuador y de relación con la seguridad vial en Ecuador como el Ministerio de Transporte y Obras Públicas (MTOP), Comisión de Tránsito del Ecuador (CTE), Gobiernos Autónomos Descentralizados Municipales (GADMs) y la Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Vial (ANT) han aprovechado los avances tecnológicos para recolectar información sobre los siniestros viales, generando grandes volúmenes de datos. Esto ha permitido cuantificar la tasa de mortalidad y sus causas más frecuentes. Así mismo, la evolución de la inteligencia artificial y la minería de datos permiten la explotación y procesamiento de las diferentes fuentes de datos, aprovechando las capacidades de cómputo para generar análisis en tiempos eficientes, de forma que se pueda agrupar,

caracterizar e identificar los patrones de mayor incidencia que contribuyen a los siniestros viales en Ecuador.

El presente trabajo de investigación aplicada se centra en el uso del aprendizaje automático y los métodos de agrupamiento no supervisados, de forma que permita identificar los conglomerados con mayor incidencia de siniestros viales (subconjuntos similares entre sí) dentro del conjunto de datos recolectados por las entidades públicas.

El análisis determinará e identificará los conglomerados y posteriormente se estudiará los principales patrones o relaciones de las causas más comunes de los siniestros viales, de ahí la importancia del estudio ya que permitirá generar acciones acordes con la realidad, focalizando estrategias de acción en los grupos identificados por las entidades gubernamentales encargadas de la seguridad vial.

A partir de este estudio se pretende dar a conocer las incidencias más comunes de siniestralidad vial. Por lo tanto, el trabajo de investigación aplicada brinda la oportunidad de profundizar nuevas aristas sobre la siniestralidad vial y las principales características y clústeres, además de poder contribuir como fuente de información para el desarrollo de programas o políticas adecuadas en la prevención de accidentes de tránsito.

1.2. Justificación

Según las Naciones Unidas (2020), la mortalidad en las vías y el mundo ascienden a 1.35 millones de personas al año, siendo la 1ra causa de muerte en niños de 5 a 14 años y adultos de 15 a 29 años y 8va causa de muerte en general por encima de enfermedades comunes como la malaria, tuberculosis y sida. El 54% de los fallecidos son personas consideras vulnerables como peatones, ciclistas y motociclistas. (Naciones Unidad, 2020)

Las estadísticas ubican a Ecuador como el 5to país de América Latina con mayor mortalidad en sinestros viales es imperioso analizar las principales características de siniestros viales en Ecuador de forma que se pueda evitar los decesos o lesiones. Toda muerte por siniestros viales es prevenible si se cuenta con leyes, políticas, herramientas

e información adecuada, de manera que se puedan generar estrategias adecuadas en la prevención de sinestros viales.

Además, los siniestros viales originan pérdidas económicas, considerando no solo el fallecimiento, sino también las lesiones que se generan impidiendo de trabajar temporal o permanentemente a una persona, esto repercute económicamente al núcleo familiar, además de los costos médicos y de rehabilitación.

De la misma forma, las altas inversiones en seguridad vial afectan el ámbito social como educación, salud, trabajo, protección social, obras e infraestructura, ya que el gobierno reduce el Presupuesto General del Estado, en busca de soluciones a los siniestros viales considerados como un problema de salud pública.

Por lo tanto, al proveer un análisis adecuado sobre las agrupaciones y características de mayor incidencia en siniestros viales, permite focalizar políticas o programas de prevención de accidentes de tránsito efectivas.

1.3. Planteamiento del Problema

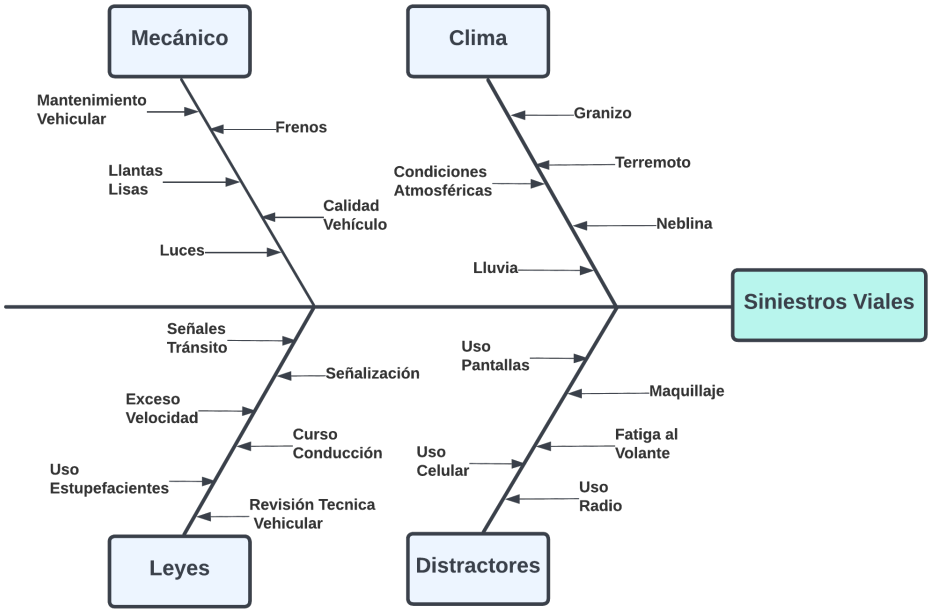
La Organización Mundial de la Salud (OMS), categoriza a los siniestros viales como un problema de salud pública que afecta negativamente el bienestar de los individuos, la población y la economía de los países, de igual forma, los accidentes de tránsito originan numerosos tipos de invalidez a los usuarios considerados vulnerables (peatones, ciclistas y motociclistas). Estas lesiones tienen gran peso en los centros de rehabilitación, así como importantes repercusiones económicas, de los costos directos, como la atención médica de los accidentados, gastos correspondientes a servicios administrativos, indemnización a vehículos, caminos y propiedades, deben sumarse los costó indirectos, como la pérdida de productividad de las víctimas y el incremento de pobreza por perdida de trabajo (Bangdiwala & Anzola, 1987).

De igual forma, el Banco Mundial (2020) ubican a Ecuador como el 5to país de América Latina con mayor mortalidad en sinestros viales. En cifras, la Agencia Nacional de Transito del Ecuador registró 21.352 siniestros de tránsito en el 2021 cuya tasa de mortalidad por cada 100.000 habitantes es de 12, lo cual representa un

incremento de 2.9 puntos con respecto al 2020 (Agencia Nacional de Transito del Ecuador, 2021).

El siguiente diagrama de causa efecto muestra los principales factores involucrados en los siniestros viales.

Figura 1-3 Diagrama de Causa y Efecto, Siniestros Viales



Nota: Diagrama de causa y efecto de siniestros viales. Fuente Maza. E, 2022

1.4. Objetivos de la Investigación

1.4.1. Objetivo General

Aplicar técnicas de agrupamiento para identificar clústeres y patrones que mayormente contribuyen a los casos de siniestralidad vial en Ecuador, de forma que se logre caracterizar los eventos más recurrentes, proporcionando información relevante a los organismos de control vial para que se promuevan programas o políticas adecuadas en la prevención de accidentes de tránsito en Ecuador.

1.4.2. Objetivos Específicos

a. Identificar los clústeres con características similares de siniestros viales y proporcionar descripciones confiables a dichas agrupaciones.

b. Mostrar que los algoritmos de agrupamiento ayudan a identificar grupos y patrones sin el conocimiento previo de la información.
c. Recomendar y focalizar los aspectos principales que deben incluir los programas o políticas de prevención de accidentes de tránsito en Ecuador.

1.5. Alcance del Trabajo

En el presente trabajo de titulación, Aplicación de técnicas de agrupamiento para caracterizar patrones de siniestros viales en Ecuador en el año 2021 se utilizará técnicas de aprendizaje automático, específicamente métodos no supervisados, cuyos algoritmos basan su proceso de entrenamiento en conjuntos de datos no etiquetados o clases previamente definidas por el científico de datos. Con estas herramientas, se busca satisfacer los objetivos de esta investigación aplicada, encontrando clústeres y características de las incidencias más comunes en los siniestros viales causados en el Ecuador en el año 2021.

De la misma forma, se considera la revisión de trabajos científicos-técnicos relacionados al tema que permitan utilizar los métodos no supervisados adecuados, que permita afinar los resultados sobre los clústeres y características que más inciden en los siniestros viales del Ecuador en el año 2021

Igualmente, para el desarrollo del trabajo de titulación y la aplicación de aprendizaje automático se adoptará la metodología CRIPS DM, cuyo modelo estándar de minería de datos permite estructurar objetivamente un proyecto, desde la etapa de compresión del problema o entendimiento del negocio hasta el lanzamiento productivo de sistemas automatizados analíticos, predictivos y/o prospectivos.

Capitulo 2

2. Marco Teórico

En este capítulo se profundizan los conceptos más relevantes sobre aprendizaje automático, clasificación no supervisada, medidas de rendimiento y medidas de validación cuyas definiciones apoyan la presente investigación aplicada. Así mismo, se expone la metodología CRIPS DM, que permite el desarrollo y descubrimiento de patrones mediante la adopción metodológica y estándar de minería de datos.

2.1. Machine Learning

Machine Learning o Aprendizaje Automático es parte de las ciencias computacionales y un subcampo de la Inteligencia Artificial que ha ido evolucionando en el tiempo, desarrollando y afinando algoritmos computacionales, cuyo fin es la emulación la inteligencia humana. El aprendizaje automático se consideran un campo en auge y de mucha importancia para extraer información y conocimiento de grandes volúmenes de datos, como en el campo de Big Data.

Los modelos y técnicas fundamentales del aprendizaje automático se están aplicado en varios campos como: las finanzas, la biología computacional, las aplicaciones biomédicas y médicas, el reconocimiento de patrones, pronósticos meteorológicos, recomendación y ventas de productos. Estas áreas de aplicabilidad y complejidad han permitido desarrollar modelos involucrando varias etapas de entrenamiento, pruebas, validaciones y afinamiento, obteniendo como resultado algoritmos de aprendizaje automático apropiados, soportando e impulsando la automatización y optimización de procesos empresariales.

En fin, que los algoritmos de aprendizaje automático puedan aprender de datos históricos o actuales y generalizar el conocimiento adquirido en tareas similares o nuevas, de modo que permita mejorar los modelos, así como la seguridad de su aplicación en áreas críticas.

2.1.1. Categorías del Machine Learning

En función de su utilidad y aplicabilidad los algoritmos de Machine Learning se dividen en 3 categorías, siendo las dos primeras las más comunes: aprendizaje supervisado y aprendizaje no supervisado y, por último, el aprendizaje por refuerzo.

Aprendizaje Supervisado: Este algoritmo requiere conjuntos de datos previamente etiquetados para entrenar los modelos y encontrar la solución esperada, este tipo de aprendizaje es aplicable a la clasificación supervisada en general. (Alhojely, 2016).

Aprendizaje No Supervisado: Este algoritmo, a diferencia de los algoritmos supervisados no dispone de datos etiquetados para el entrenamiento. Para estos algoritmos, solo se conoce son los datos de entrada y su finalidad radica en encontrar patrones en los datos que permita simplificar el análisis. Los algoritmos más utilizados son: Algoritmos de Clústeres, Análisis de Componentes Principales y la Detección de Anomalías.

Aprendizaje por refuerzo (Reinforcement): Este algoritmo, a diferencia del aprendizaje supervisado y no supervisado, fundamenta su entrenamiento en recompensas y castigos, a medida que se resuelve el problema, este modelo trata de formular la mejor estrategia para obtener el mayor rendimiento en el tiempo.

2.1.2. Aprendizaje no Supervisada

Los algoritmos y técnicas de clústeres no supervisada son ampliamente usadas en el ámbito de la ciencia de datos. La clasificación no supervisada, se refiere esencialmente a la colección de algoritmos o métodos (estadísticos y no estadísticos) que permiten agrupar elementos de un conjunto de datos con características similares, sobre los cuales se aplican diferentes métricas de validación y rendimiento. Los elementos que revelen características similares entre sí quedan agrupados en conjuntos que llamaremos clústeres. Estos clústeres se irán formando en base a las características de los elementos, métricas y validaciones del algoritmo no supervisado. La literatura elemental y científica sobre clasificación no supervisada es numerosa, algunas fuentes secundarias como (Ghahramani, 2003; Celebi y Aydin, 2016; Sinagay Yang, 2020).

Cabe destacar que la aplicación de técnicas de clasificación no supervisa en conjuntos de datos no tiene una prescripción ideal para descubrir patrones o que sean formulas universalmente aplicables para descubrir estructuras, relaciones o clústeres que pueden estar presentes en datos multidimensionales. Así mismo, no todos los algoritmos de clasificación no supervisada pueden manifestar los agrupamientos presentes en los datos, dado que estos algoritmos hacen suposiciones implícitas acerca de la forma de los agrupamientos, basándose en medidas de similaridad.

Entre las principales características del aprendizaje no supervisado se destacan las siguientes:

- Comúnmente utilizados para encontrar información útil a partir de los datos.
- Su aprendizaje es equivalente a los seres humanos, aprende en base a la información que extrae de los datos.
- Funciona con datos sin etiquetas, ni categorías.
- En el mundo real es complejo obtener datos etiquetados y que faciliten los casos de estudio, de ahí la importancia del aprendizaje no supervisado.

Finalmente, podemos hablar de varios tipos de algoritmos de clústeres en aprendizaje no supervisada como: Algoritmo de agrupamiento K-means, basado en centroides. Algoritmo de clústeres DBSCAN, basado en densidad local de aplicaciones con ruido disperso. Algoritmo de agrupamiento de Jerarquía Aglomerativa, basado en dendrogramas.

2.1.3. Métodos de Agrupación

En esta sección, proporcionamos los detalles de los algoritmos de agrupación K-means, Jerárquico Aglomerativa y DBSCAN que se han presentado brevemente en la Sección 2.1.2.

2.1.3.1. K-means

La agrupación de K-means es uno de los algoritmos de aprendizaje no supervisado más sencillos y utilizados, especialmente en minería de datos y estadística. Al tratarse de un algoritmo de partición, su objetivo es formar grupos de puntos de datos basados en el número de conglomerados, representados por la variable k. El valor de k suele

ser desconocido a priori y debe ser elegido por el científico de datos utilizando técnicas como el método de Elbow. Cada conglomerado tiene un centroide, que suele calcularse como la media de los vectores de características del conglomerado. La pertenencia a un clúster de cada elemento de datos en el algoritmo de agrupación k-means se decide en función del centroide del clúster más cercano al punto.

El algoritmo k-means divide un conjunto de datos en k clústeres empleando los siguientes pasos:

1. Inicializa los centroides de los clústeres con k valores iniciales.
2. Para formar los k conglomerados, cada punto del conjunto de datos se asigna al centroide más cercano en función de la distancia.

La *distancia euclidiana* se utiliza para calcular la distancia entre cada punto de datos y los centroides inicializados. Existen otras métricas para encontrar la distancia más cercana, aplicamos la distancia euclídea porque varias investigaciones anteriores sobre análisis de agrupación obtuvieron grandes resultados utilizando la distancia euclídea.

---

3. Los centroides se recalculan promediando todos los puntos de datos de cada conglomerado para reducir la varianza.
4. Los pasos 2 y 3 iteran hasta que se cumple algún criterio.

Los criterios son normalmente, cuando no hay cambios en los valores de los centroides, la suma de distancias entre los puntos de datos y el centroide de cada clúster ya no cambia, los puntos de datos asignados a los clústeres son los mismos que la asignación anterior o se ha alcanzado el número máximo de iteraciones en el caso de que el algoritmo tenga tiempos de iteración fijos.

Matemáticamente la distancia euclidiana entre dos puntos de datos X_1 y X_2, cada uno representado por un vector $p - dimensional$, $X_1 = (X_{1_1}, X_{1_2}, \ldots, X_{1_p})$ y $X_2 = (X_{2_1}, X_{2_2}, \ldots, X_{2_p})$, se denota como $d\_Euc\ (X_1,\ X_2)$, y se define como sigue:

$$d\_Euc\ (X_1, X_2) = \sqrt{\sum_{i=1}^{p} (X_{1_i} - X_{2_i})^2}$$

El procedimiento de agrupación k-means desplaza iterativamente los puntos de datos entre clústeres, minimizando la suma de distancias al cuadrado, denotada por J,

de cada punto de datos a su centroide de clúster. Denotemos el i^{th} clúster por C_i, entonces la suma de distancias al cuadrado para C_i, denotada por J, se define como sigue:

$$J_i = \sum_{X \in C_i} d\_Euc(X, Y)^2$$

donde $d\_Euc(X, Y_i)$ es la distancia euclídea de un punto de datos X en C_i a $C_i's$ al centroide Y_i.

A continuación, podemos calcular la suma de distancias al cuadrado para todos los k clústeres, denotada por J, como:

$$J = \sum_{i=1}^{k} J_i$$

Iniciar el algoritmo k-means a partir de varios conjuntos de valores iniciales puede dar lugar a distintos mínimos locales de J. Queremos encontrar el que sea el mínimo global. Sin embargo, no es realista agotar todos los conjuntos de valores iniciales. Por lo tanto, ejecutamos la agrupación k-means varias veces, partiendo de diferentes valores iniciales en cada ejecución y elegimos la solución que minimiza la suma de distancias al cuadrado, J. Mediante el uso de múltiples ejecuciones, es más probable que el algoritmo converja al mínimo global de J, o al menos a un mínimo local que sea el más cercano al mínimo global entre los múltiples mínimos locales.

Ventajas

- Dado que k-means es un algoritmo de agrupación simple, puede implementarse fácilmente.
- K-means compara la distancia entre los puntos de datos y agrupa los clústeres. Por lo tanto, puede ser computacionalmente más rápido que la agrupación jerárquica.

Desventajas

- El número de conglomerados k, deben especificarse manualmente.

- Los resultados de la agrupación pueden variar en función de los valores iniciales. K-means también selecciona aleatoriamente los centroides iniciales para k clústeres. Por lo tanto, los resultados pueden ser diferentes de una ejecución a otra.
- K-means tiene dificultades para agrupar conjuntos de datos de tamaño y densidad variables.
- K-means no puede identificar valores atípicos.

2.1.3.2. Agrupamiento Jerárquico

Los algoritmos de agrupamiento jerárquico buscan construir una jerarquía de clústeres. Comienza con algunos conglomerados iniciales y converge gradualmente hacia la solución. La agrupación jerárquica tiene dos categorías: aglomerativo y divisivo. El enfoque aglomerativo toma inicialmente cada punto de datos como un conglomerado individual e iterativamente fusiona los conglomerados hasta que el conglomerado final contiene todos los puntos de datos. Como técnica opuesta al agrupamiento aglomerativo, la técnica de agrupación jerárquica divisivo sigue un flujo descendente que parte de un único clúster que contiene todos los puntos de datos y lo divide iterativamente en clústeres más pequeños hasta que cada clúster contiene un punto de datos. En este trabajo de investigación aplicada, utilizamos un enfoque aglomerativo para agrupar el conjunto de datos.

El algoritmo de agrupación jerárquica aglomerativo tiene los siguientes pasos:

1. El algoritmo toma cada punto de datos como un único clúster y decidimos una matriz de proximidad específica para determinar la distancia entre los clústeres.

Hay cuatro funciones de distancia para la matriz de proximidad: single linkage (min), average linkage, complete linkage y ward linkage (max). Single linkage significa que la distancia entre dos clústeres se define como la distancia mínima entre un punto del primer clúster y otro punto del segundo clúster. Complete linkage toma como distancia entre dos conglomerados el valor máximo de la distancia entre dos puntos de datos. Average linkage calcula la distancia de todos los puntos de datos del primer conglomerado con todos los demás del segundo conglomerado y toma la distancia media como distancia entre los conglomerados. Ward linkage es similar a average linkage excepto en que utiliza la suma de cuadrados para calcular la distancia entre los puntos. En este trabajo de investigación aplicada, utilizamos Ward linkage como función de distancia.

---

2. Para encontrar el par de clústeres más cercano, calcula la similitud (distancia) entre cada uno de los clústeres.
3. Los conglomerados similares se fusionan para formar un conglomerado acorde a la función de distancia.
4. Se iteran los pasos 2 y 3 hasta que todos los puntos de datos se fusionan en un último conglomerado.

En fin, la agrupación jerárquica forma un único árbol de conglomerados en el que cada nodo representa los conglomerados y cada punto de datos comienza como una hoja del árbol.

Ventajas

- No se requiere especificar el número de clúster.
- Al igual que k-means, los algoritmos de agrupamiento jerárquico son fáciles de implementar.
- Pueden mostrar la estructura jerárquica de un árbol de conglomerados (dendrogramas), lo que puede ayudar a decidir el número de conglomerados.

Desventajas

- No hay retrocesos, lo que significa que, una vez creado un clúster, los puntos de datos de pertenencia no pueden moverse.
- Dependiendo de la elección de la matriz de distancias, puede ser sensible a ruidos y valores atípicos. Además, puede tener dificultades para manejar conglomerados de diferentes tamaños y formas convexas.

2.1.3.3. DBSCAN

El algoritmo DBSCAN no requiere la especificación a priori del número de clústeres. Sin embargo, DBSCAN requiere dos parámetros: eps y min_sample. Eps define la distancia máxima entre dos muestras de datos en la que se supone que una de ellas es próxima de otra que es un punto central de un clúster. min_sample define el número mínimo de muestras que deben estar en la vecindad junto con la muestra núcleo. Se asume que un clúster es una región densa con puntos de datos que es mayor

que min_sample dentro del rango de eps del punto central y cada clúster está separado de otro por una densidad menor.

El algoritmo DBSCAN define los siguientes pasos:

1. Comienza con un punto de datos inicial arbitrario que no ha sido visitado. La vecindad de este punto se extrae utilizando la máxima distancia eps.
2. Si su vecindad contiene el número suficiente de puntos según min_sample, comienza la agrupación. Ese punto inicial se convierte también en el punto central del clúster. En caso contrario, el punto se considera ruido. Más adelante, es probable que este punto se encuentre en la vecindad de otro punto y, por tanto, forme parte de un conglomerado. En cualquier caso, este punto se marca como punto visitado.
3. A continuación, los puntos de la vecindad del punto núcleo se utilizan para buscar sus respectivos puntos de vecindad, ya que estos puntos siguen siendo nuevos puntos no visitados.
4. Se repiten pasos 2 y 3 hasta que todos los puntos del clúster hayan sido visitados y etiquetados formando el clúster conectado por densidad.
5. A continuación, se recupera el nuevo punto no visitado del conjunto de datos y el algoritmo repite los pasos 1 al 4 hasta que todos los puntos hayan sido visitados y pasen a ser ruido o parte de un conglomerado.

Ventajas

- DBSCAN funciona bien con el conjunto de datos que insiste en los clústeres de alta densidad frente a los de baja densidad.
- Es resistente al ruido y puede manejar valores atípicos del conjunto de datos.
- Manejar clústeres de diferentes formas y tamaños.
- A diferencia de k-means, no es necesario definir a priori el número de conglomerados.

Aunque DBSCAN tiene la ventaja de no requerir un valor predefinido para el número de clústeres, el valor de la distancia, eps, resulta difícil de estimar. Especialmente cuando los clústeres tienen una densidad variable.

Desventajas

- DBSCAN puede separar los conglomerados de alta densidad de los de baja densidad, no funciona bien con conglomerados de densidades variables o de densidad similar.
- BDSCAN no puede manejar bien datos de alta dimensión.

2.1.4. Hiperparámetros y Parámetros

2.1.4.1. K-means y el método de Elbow

La eficiencia del algoritmo K-means depende principalmente de encontrar el valor de k optimo. El método más utilizado para encontrar el numero óptimo de clústeres se denomina Elbow (codo), que utiliza el concepto de valor WCSS (Within Cluster Sum of Squares), este calcula la varianza de las distancias totales entre cada punto de datos y su centroide.

El algoritmo determina los valores WCSS para diferentes valores de k en el rango de 1 a 10 usualmente. Traza una curva entre los valores calculados y el número de clústeres de k, el punto afilado o codo es el mejor valor de k.

2.1.5. DBSCAN y los parámetros eps y $minPts$

DBSCAN es un algoritmo de agrupamiento basado en densidad de aplicaciones con ruido, los parámetros fundamentales son:

- eps: Especifica la distancia de los vecindarios. Dos puntos se consideran vecinos si la distancia entre ellos es menor o igual a eps.
- $minPts$: número mínimo de puntos para definir un clúster.

En base a los parámetros eps y $minPts$, los puntos pueden categorizar como punto central, punto fronterizo (borde) y valor atípico:

- ***Punto central:*** si hay al menos un número mínimo de puntos ($minPts$) en su área circundante con radio eps.
- ***Punto de borde:*** si es accesible desde un punto central y hay menos de $minPts$ de puntos dentro de un radio de eps a su alrededor.

- ***Valor atípico:*** un punto es un valor atípico si no es un punto central y no es accesible desde ningún punto central.

Por lo general $minPts$ debe ser mayor o igual al número de dimensiones más uno. Así mismo, si el conjunto de datos tiene valores atípicos $minPts$ debe ser un número grande, dado que elimina los puntos de ruido fácilmente.

Para determinar el valor óptimo de eps se puede utilizar el método de Elbow, como:

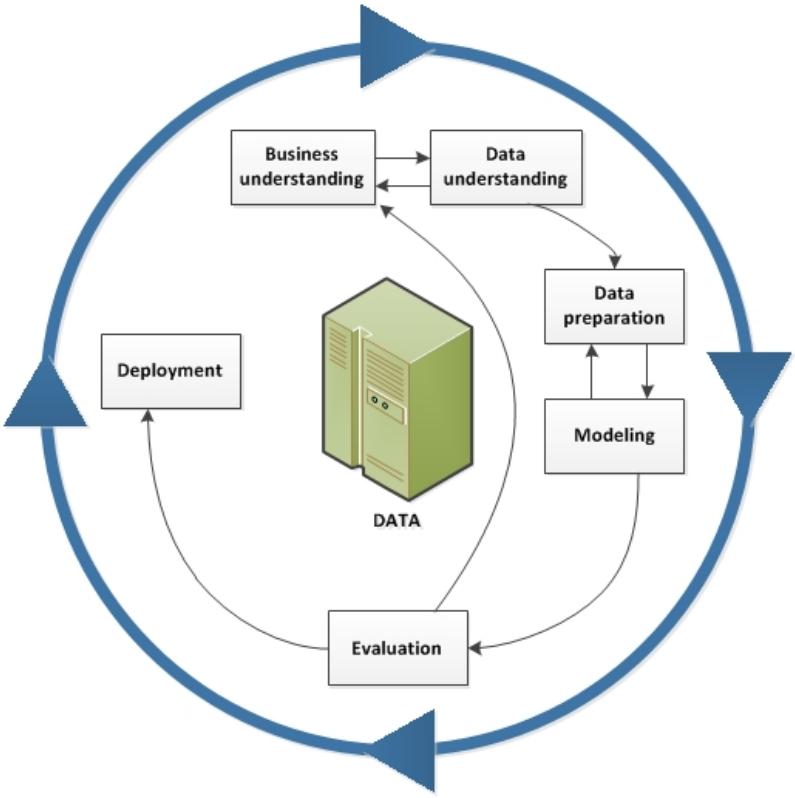
- Calculamos la distancia d para cada punto de datos x y ordenamos las distancias en orden ascendente.
- Trazamos un gráfico entre la distancia y el índice de puntos
- Elegimos la distancia optima eps, donde exista un fuerte incremento.

2.2. CRIPS DM

El desarrollo de la presente investigación aplicada requiere la adopción de un marco o metodología que proporcione un enfoque estándar para la extracción de patrones o características de los conjuntos de datos, por eso se adoptado la metodología CRIPS DM.

CRIPS DM – Cross Industry Standard Process for Data Mining, es un modelo abierto que describe las fases de un proyecto de minería de datos y que es utilizado ampliamente en el campo de la ciencia de datos ya que permite alinear cualquier tema de interés o estudio asociado con la minería de datos. El modelo de DRIPS DM define seis fases en el desarrollo de un ejercicio de minería de datos tal y como se visualiza en la Figura 4. Estas fases corresponden a la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y finalmente el despliegue.

Figura 2-1 Fase modelo CRIPS DM



Nota. Metodología CRIPS DM para proyectos de minería de datos. Tomado de (Metodología CRIPS DM, 2021)

A pesar de que el modelo CRIPS DM define el flujo de las fases, la sucesión de estas no es necesariamente estricta, es decir se permite la iteración recursiva entre las diferentes etapas. Por otro lado, la recurrencia que admite CRIPS DM posibilita afinar cada etapa dentro de la ejecución del proyecto, de hecho, una vez obtenidos los resultados, se puede repetir las fases hasta obtener los resultados adecuados en el proyecto de minería de datos. Cada una de las fases del modelo CRIPS DM se describen a continuación:

a. ***Comprensión del Negocio*** (Business Understanding)

Fase inicial que asimila los objetivos y los requerimientos desde una perspectiva de interés para el negocio. Analiza la situación actual, y traduce los objetivos del negocio a términos de un problema de minería de datos, definiendo las metas de la minería de datos, además en esta fase se diseña un plan preliminar para el desarrollo del proyecto.

b. ***Comprensión de Datos*** (Data Understanding)

Fase que involucra el proceso de recopilación y exploración de datos iniciales. A partir de estas tareas se realiza la descripción de los datos y se examinan características de cantidad de datos, tipos de valores, codificación de los datos. Así mismo, se realiza la verificación de la calidad considerando los datos perdidos, metadatos erróneos, incoherencia de codificación. Finalmente se genera un resumen que puede incluir estadísticas básicas, tablas, visualizaciones e inferencias que permiten suscitar premisas o hipótesis iniciales.

c. ***Preparación de Datos*** (Data Preparation)

Fase que involucra los aspectos más importantes de la minería de datos y con frecuencia las tareas relacionadas toman en promedio el 50 – 70 % del tiempo y esfuerzo de un proyecto de ciencia de datos. Dedicar el tiempo adecuado en las fases anteriores puede reducir la complejidad e incluso los gastos indirectos. La preparación de los datos usualmente se realiza varias veces y sin un orden prescrito, frecuentemente implica tareas como agregación de registros, formateo de datos, integración de conjuntos, derivación de nuevos atributos, clasificación de datos, eliminación o sustitución de valores.

d. ***Modelado*** (Modeling)

Fase que involucra el desarrollo y la selección del modelo que cumpla con la compresión del negocio y sus objetivos. El modelado generalmente se ejecuta en múltiples iteraciones. Los científicos de datos u analistas de datos prueban diferentes algoritmos utilizando los parámetros predeterminados y posteriormente se ajustan los parámetros a valores óptimos. Asimismo, dependiendo del resultado, de las métricas de evaluación y comportamiento del modelo, es probable que se regrese fase anterior para modificar la selección de variables o ajustar la limpieza de datos.

e. ***Evaluación*** (Evaluation)

Fase que permite determinar si los modelos son técnicamente correctos y eficientes en base a los criterios establecidos en las fases anteriores, de la misma manera, se debe evaluar los resultados obtenidos en función de los criterios establecidos al inicio del proyecto, por lo tanto, en necesario asegurar que la organización o negocio puede utilizar el modelo y los resultados obtenidos que generalmente reciben el nombre de *descubrimientos*. Por último, es importante realizar una evaluación del proceso realizado, interpretar los resultados y verificar que el modelo cumple objetivos planteados.

f. ***Despliegue*** (Deployment)

Fase que disponibiliza el uso del conocimiento adquirido para implementar mejoras en la organización u negocio. En general esta fase incluye las actividades de planificación y control del despliegue de resultados, donde se determina el plan de despliegue y los requisitos base. Además, se documenta las actividades o tareas realizas en el proyecto de minería de datos. En función de los requerimientos de la fase de despliegue también se define los pasos de despliegue, la integración con otros sistemas, requisitos base de hardware y software, plan de contingencia, solución a problemas comunes y la documentación técnica relacionada al proyecto de minería de datos.

Capitulo 3

En este capítulo se cubrirá el desarrollo de la investigación aplica, se buscará responder a los objetivos, mediante la aplicación de métodos de agrupamiento de aprendizaje no supervisado, se utilizará la metodología de CRISP DM, que permite procesar de forma ordenada el conjunto de datos obtenidas de fuentes secundarios. Además, se revisará literatura de trabajos investigación que aporten al presente trabajo de investigación aplicada.

3. Metodología

3.1. Desarrollo de la Metodología

La metodología planteada CRISP DM es un marco de trabajo estandarizado para proyectos de minería de datos, por esta razón se adopta esta metodología para aplicar técnicas de agrupamiento en el conjunto de datos para caracterizar patrones de siniestros viales en Ecuador en el año 2021. Por lo tanto, se adoptan las fases y tareas que contribuyen al desarrollo de los objetivos.

3.1.1. Compresión del Tema de Interés

Como se planteó en el Capítulo 1 el propósito del presente trabajo de investigación aplicada es identificar clústeres y patrones que mayormente contribuyen a los casos de siniestralidad vial en Ecuador, de forma que se logre caracterizar los eventos más recurrentes, proporcionando información relevante a los organismos de control vial para que se promuevan programas o políticas adecuadas en la prevención de accidentes de tránsito en Ecuador.

3.1.2. Compresión de los Datos

Antes de iniciar con la preparación de los datos, es importante acceder a los datos primarios o datos sin procesar que tenemos de los siniestros viales en el Ecuador en el año 2021 y explorarlos con ayuda de tablas, gráficos que permitan determinar la calidad de datos a priori obtenidos de las diferentes fuentes, inclusive tener un entendimiento inicial de sus características.

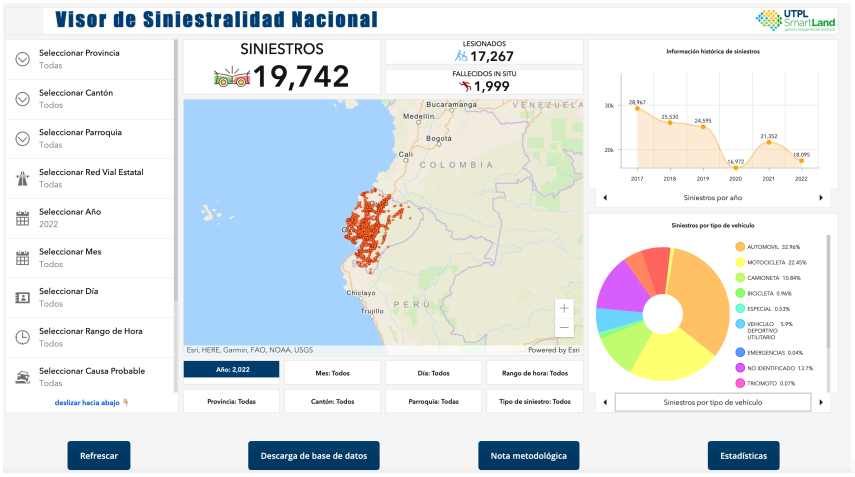
3.1.2.1. Recopilación de Datos Iniciales

Para el año 2021 las entidades públicas del Ecuador y de relación con la seguridad vial en Ecuador como el Ministerio de Transporte y Obras Públicas (MTOP), Comisión de Tránsito del Ecuador (CTE), Gobiernos Autónomos Descentralizados Municipales (GADMs) y la Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Vial (ANT) cuentan con diversas fuentes de información sobre los siniestros viales, entre las cuales se pueden encontrar bases de datos con los siguientes aspectos:

Base de Datos de Siniestros Viales

- La página web de la Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Vial de Ecuador, reúne información de varios aspectos como: movilidad, siniestros viales, control de transporte y seguridad vial, este trabajo se enfoca en las bases de datos del componente de siniestralidad vial.
- Para obtener la información, se debe ingresar al sito de web https://www.ant.gob.ec/ y utilizar la opción descarga como se muestra en la figura 5. Una vez realizando este proceso, el sitio descarga un archivo en formato *cvs* que puede ser leído en Excel o directamente sobre las herramientas de minería de datos.

Figura 3-1 Recopilación de datos de siniestros viales en el año 2021 del Ecuador



Nota. Estadísticas de siniestralidad vial del Ecuador en el año 2021. Tomado de (Agencia Nacional de Tránsito, 2021)

3.1.2.2. Descripción de Datos

En línea con la compresión del tema de interés y el objetivo del presente tema de investigación aplicada, en este punto se inicia con la descripción y exploración de las bases de datos que contienen los registros sobre los siniestros viales ocurridos en Ecuador en el año 2021.

- Base de Datos de Siniestros Viales

Esta base de datos se compone de 21.352 registros y 56 columnas que pueden organizarse en las siguientes categorías como se muestra en la Tabla 3-1.

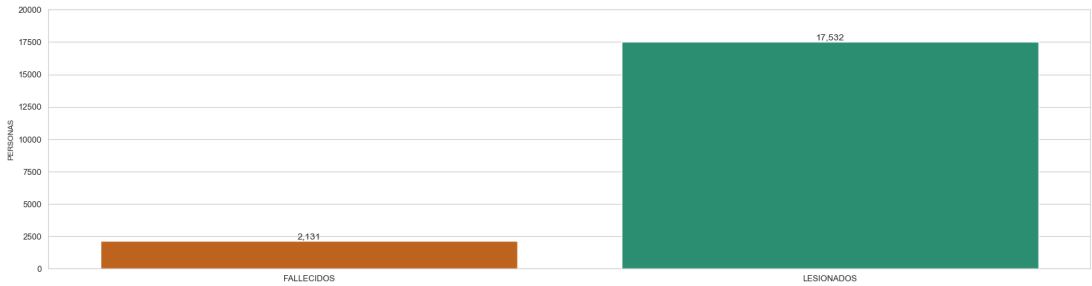
Tabla 3-1 Campos que componen la base de datos de sinestros viales del Ecuador en el año 2021

Categoría	Campos
Temporal	Fecha del siniestro, hora
Espacial	Latitud, longitud, provincia, cantón, parroquia, dirección
Demográfica	Sexo, edad
Información del siniestro	Tipo de siniestro, causa probable, tipo de vehículo, condición.

- Información Temporal

Para la fecha en la que se obtuvo las bases de datos de fuentes secundarias mencionadas en la Recopilación de Datos Iniciales, se encuentran los datos consolidados para el año 2021 por las entidades de control. Asimismo, como se muestra en la Figura 3-2 de los sinestros registrados se reporta que hubo 2,131 personas fallecidas y 17,532 personas con diferentes tipos de lesiones.

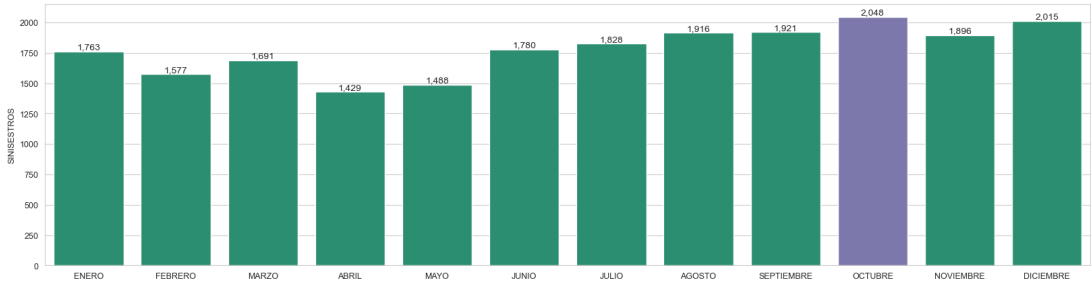
Figura 3-2 Número de personas relacionadas en siniestros viales en el año 2021 en Ecuador



Fuente (Maza. E, 2022)

Asimismo, el comportamiento mensual de sinestros viales se puede observar en la Figura 3-3 que octubre es el mes con la mayor incidencia, el cual registró 2,048 accidentes de tránsito.

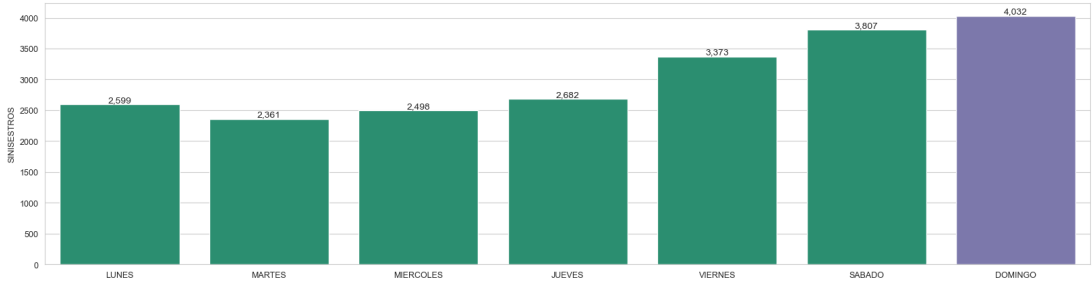
Figura 3-3 Número de siniestros viales registrados por mes en el año 2021



Fuente (Maza. E, 2022)

Del mismo modo, el comportamiento semanal de siniestros viales se puede observar en la Figura 3-4 que el domingo es el día de mayor incidencia, el cual registró 4,032 accidentes de tránsito.

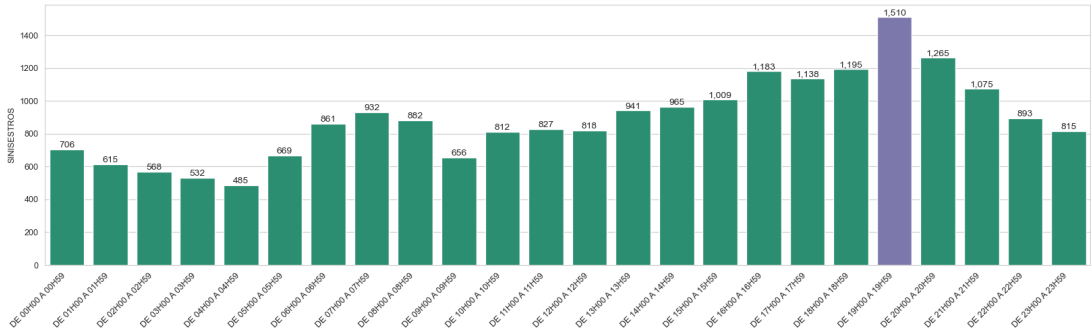
Figura 3-4 Número de siniestros viales registrados por semana en el año 2021



Fuente (Maza. E, 2022)

Igualmente, el comportamiento por periodo en horas de siniestros viales se puede observar en la Figura 3-5 que el horario de 19H00 - 19H59 es de mayor incidencia, el cual registró 1,510 accidentes de tránsito.

Figura 3-5 Número de siniestros viales registrados por periodo de horas en el año 2021

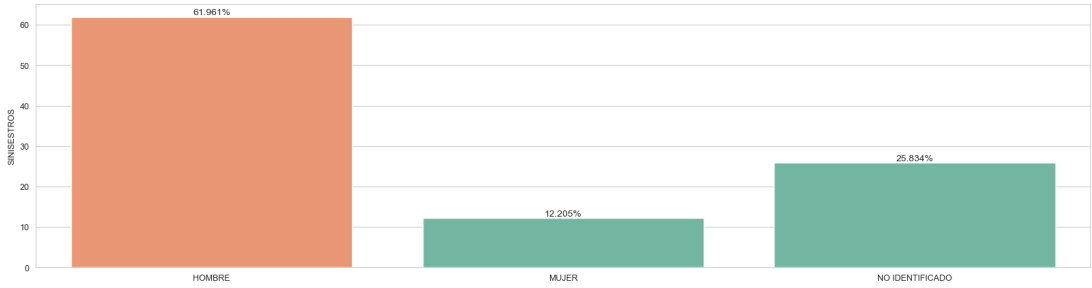


Fuente (Maza. E, 2022)

- Información Demográfica de Siniestros Viales.

La base de datos incluye información demográfica de los siniestros viales. Continuando con la descripción de los datos, se encuentra que el 61.96% de los incidentes registrados son hombres, 12.20% son mujeres y el 25.83% se reporta como no identificado.

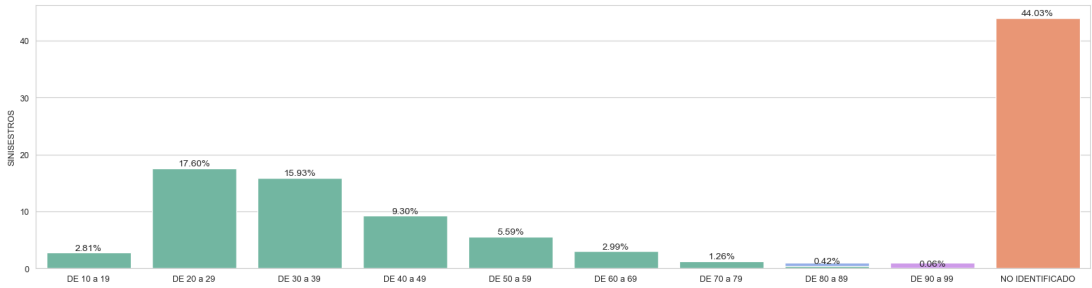
Figura 3-6 Porcentaje del género que inciden en siniestros viales registrados en el año 2021



Fuente (Maza. E, 2022)

Para la característica de edad, como se muestra en la Figura 3-7 el 44 % de sinestros viales no se registró la edad de las personas involucradas, sin embargo, el 48 % que mayormente inciden en los siniestros viales son las personas que se encuentran entre los 20 años a 59 años, siendo un aporte significativo del 17.6 % las personas que se encuentran entre los 20 años a 29 años.

Figura 3-7 Porcentaje de rangos de edad que inciden en siniestros viales registrados en el año 2021



Fuente (Maza. E, 2022)

Estas primers observaciones de datos temporales y demograficos ya pueden inferir algunas carateristicas de los clusteres objetivos en la presente investigación aplicada.

- Información del Siniestro Vial

Según los datos recolectados por las entidades de control existen varias cusas que inciden en los siniestraos viales, sin embargo, se han agrupado en 27 causas probables de accidentes viales registrados en el año 2021, la cuales se detallan en la siguiente tabla.

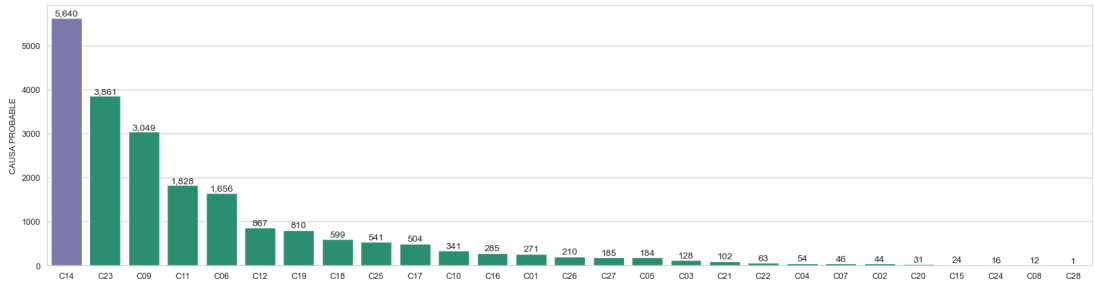
Tabla 3-2 Causas probables de siniestros viales en el Ecuador en el año 2021.

Código	Causa Probable
C01	Caso fortuito o fuerza mayor (explosion de neumatico nuevo, derrumbe, inundacion, caida de puente, arbol, presencia intempestiva e imprevista de semovientes en la via, etc.).
C02	Presencia de agentes externos en la via (agua, aceite, piedra, lastre, escombros, maderos, etc.).
C03	Conducir en estado de somnolencia o malas condiciones fisicas (suenio, cansancio y fatiga).
C04	Danios mecanicos previsibles.
C05	Falla mecanica en los sistemas y/o neumaticos (sistema de frenos, direccion, electronico o mecanico).
C06	Conduce bajo la influencia de alcohol, sustancias estupefacientes o psicotropicas y/o medicamentos.
C07	Peaton transita bajo influencia de alcohol, sustancias estupefacientes o psicotropicas y/o medicamentos.
C08	Peso y volumen - no cumplir con las normas de seguridad necesarias al transportar cargas.
C09	Conducir vehiculo superando los limites maximos de velocidad.
C10	Condiciones ambientales y/o atmosfericas (niebla, neblina, granizo, lluvia).
C11	No mantener la distancia prudencial con respecto al vehiculo que le antecede.
C12	No guardar la distancia lateral minima de seguridad entre vehiculos.
C14	Conducir desatento a las condiciones de transito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).
C15	Dejar o recoger pasajeros en lugares no permitidos.
C16	No transitar por las aceras o zonas de seguridad destinadas para el efecto.
C17	Bajarse o subirse de vehiculos en movimiento sin tomar las precauciones debidas.
C18	Conducir en sentido contrario a la via normal de circulacion.
C19	Realizar cambio brusco o indebido de carril.
C20	Mal estacionado - el conductor que detenga o estacione vehiculos en sitios o zonas que entranien peligro, tales como zona de seguridad, curvas, puentes, tuneles, pendientes.
C21	Malas condiciones de la via y/o configuracion. (iluminacion y disenio).
C22	Adelantar o rebasar a otro vehiculo en movimiento en zonas o sitios peligrosos tales como: curvas, puentes, tuneles, pendientes, etc.
C23	No respetar las seniales reglamentarias de transito. (pare, ceda el paso, luz roja del semaforo, etc).
C24	No respetar las seniales manuales del agente de transito.
C25	No ceder el derecho de via o preferencia de paso a vehiculos.
C26	No ceder el derecho de via o preferencia de paso al peaton.
C27	Peaton que cruza la calzada sin respetar la senializacion existente (semaforos o seniales manuales).
C28	Dispositivo regulador de transito en mal estado de funcionamiento (semaforo).

De acuerdo con lo expuesto en la Figura 3-8, la principal causa probable en los siniestros viales es *conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor)* cuyo aporte en accidentes de tránsito es del 26.41 % para el año 2021, las siguientes dos causas que

también inciden en los siniestros viales son *no respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc.)* con un 18.08 % y *conducir vehículo superando los límites máximos de velocidad* con un aporte del 14.08 %.

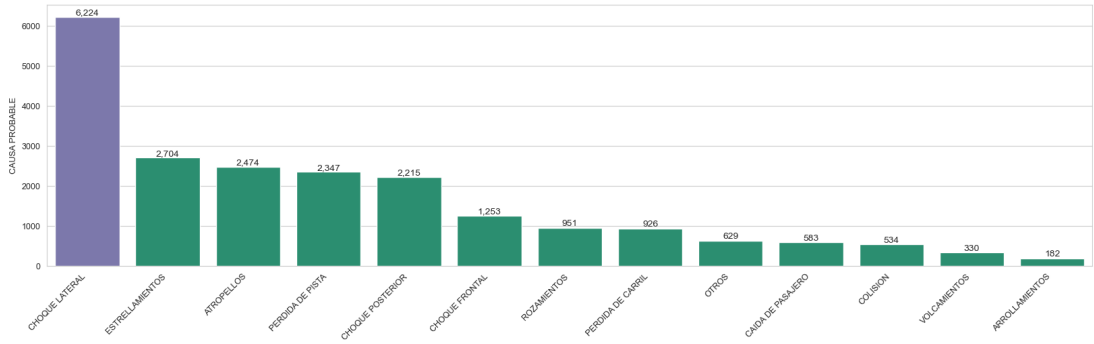
Figura 3-8 Número de siniestros viales por causas probables en el año 2021.



Fuente (Maza. E, 2022)

Igualmente, las entidades de control han clasificado el tipo de siniestro vial, se puede observar en la Figura 3-9 que el *choque lateral* es de mayor incidencia, el cual registró 6,224 accidentes de tránsito.

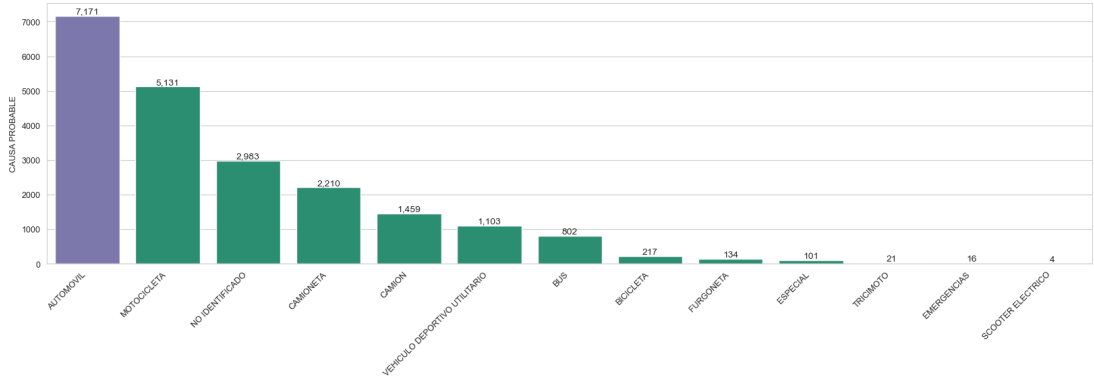
Figura 3-9 Número de siniestros viales por tipo de siniestro en el 2021.



Fuente (Maza. E, 2022)

Del mismo modo, el tipo de vehículo que más incide en los siniestros viales como se muestra en la Figura 3-10 son los automóviles, el cual registró 7,171 accidentes de tránsito.

Figura 3-10 Número de siniestros viales por tipo de vehículo en el año 2021.



Fuente (Maza. E, 2022)

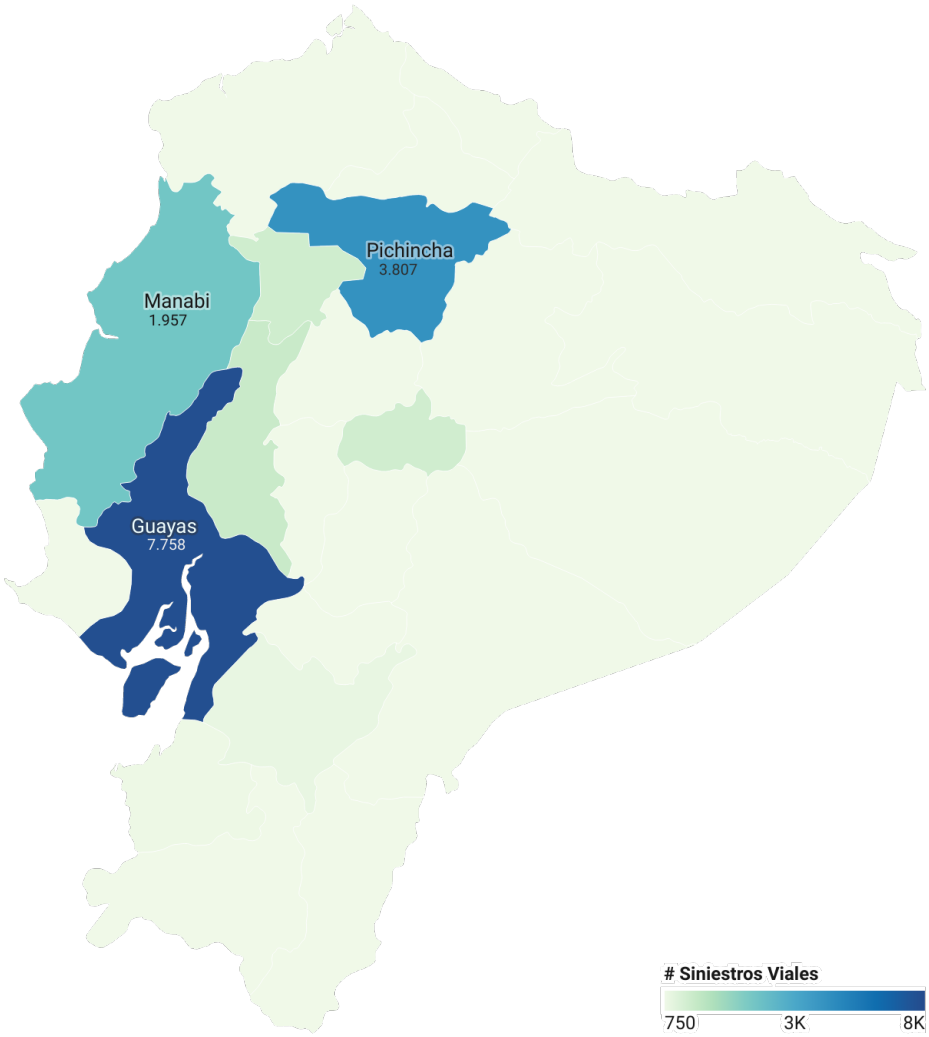
De acuerdo con las observaciones realizadas para los siniestros viales hay varias características que predominan en los accidentes de tránsito como el 26.41 % ese debe a *conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor)*, el 29.15 % son accidentes del *tipo choque lateral* y el 33.58 % de sinestros viales es caudados por *automóviles*.

Estas primeras características en los siniestros viales ya pueden inferir sobre los clústeres objetivos en la presente investigación aplicada

- Información Geográfica

Como se presenta en la Figura 3-11, la mayor incidencia de los siniestros se encuentra las provincias de Guayas, Pichincha y Manabí cuya concentración es del 63.33 %, siendo Guayas la provincia que reporto 7,758 accidentes de tránsito el año 2021.

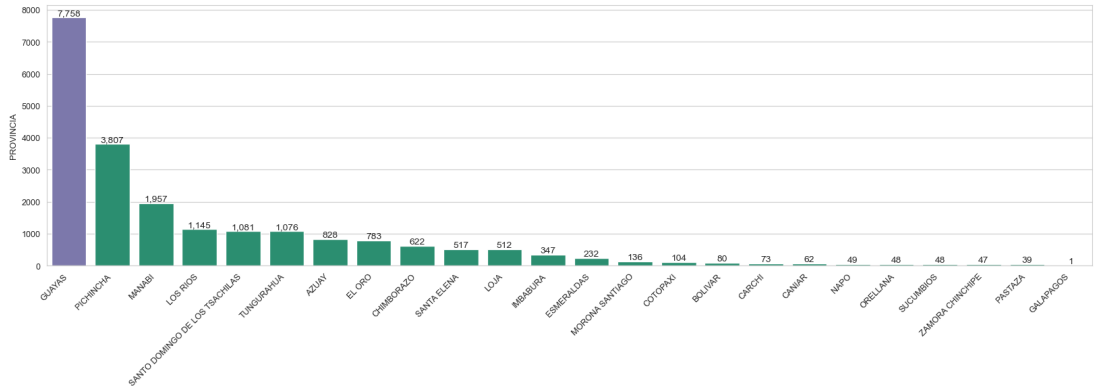
Figura 3-11 Distribución espacial de los siniestros viales por provincia del Ecuador en el año 2021.



Mapa de calor en base al número de siniestros ocurridos en Ecuador en el año 2021.
Map: Edwin Maza Jara • Source: ANT • Created with Datawrapper

Fuente (Maza. E, 2022)

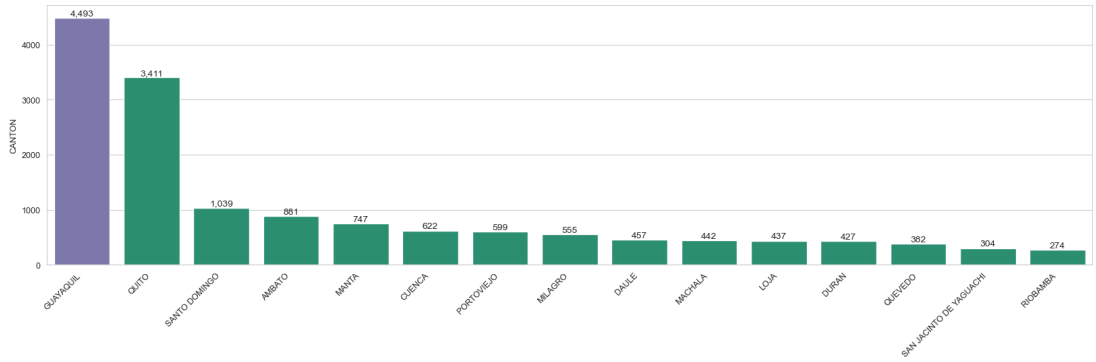
Figura 3-12 Número de siniestros viales por provincias en el año 2021



Fuente (Maza. E, 2022)

Y en una escala espacial más pequeña, se puede observar en la Figura 3-13 que el cantón de Guayaquil de la provincia Guayas, abarca el mayor porcentaje de incidentes con un 21.04 %, seguido por los cantones: Quito (15.98 %) y Santo Domingo (4.87 %).

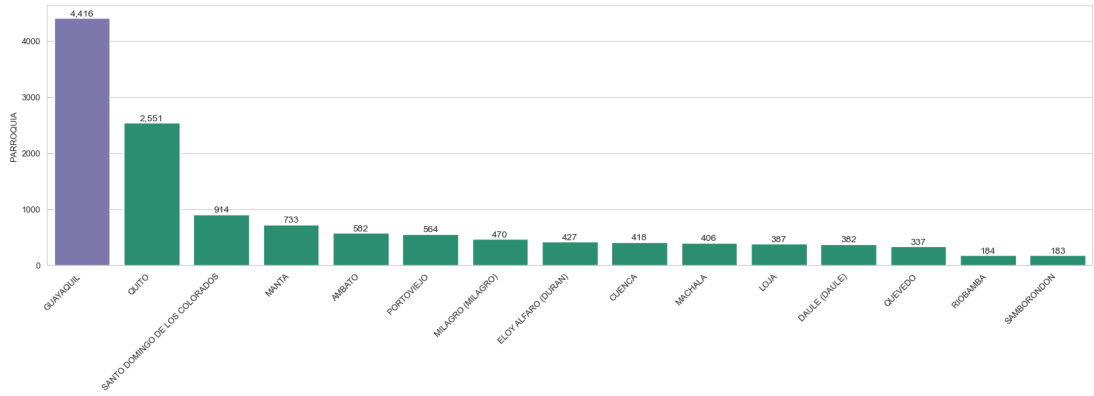
Figura 3-13 Número de siniestros viales por cantones en el año 2021.



Fuente (Maza. E, 2022)

De manera similar, en una escala espacial más pequeña, se puede observar en la Figura 3-14 que la parroquia Guayaquil de la provincia Guayas, abarca el mayor porcentaje de incidentes con un 20.68 %, seguido por las parroquias: Quito (11.95 %), Santo Domingo de los Colorados (4.28 %) y Manta (3.43 %)

Figura 3-14 Número de siniestros viales por parroquia en el año 2021.



Fuente (Maza. E, 2022)

3.1.2.3. Exploración de Datos

Como se había mencionado las entidades de control y específicamente la Agencia Nacional de Transito asocia los accidentes de tránsito en Ecuador a 27 causas definidas en la Tabla 3-2, se presenta las estadísticas descriptivas de las 10 principales causas probables que inciden en un 90.64 % de los siniestros viales registrado en Ecuador en el año 2021.

Tabla 3-3 Ocurrencia de siniestros viales por causas probables en el año 2021

Código	Causa Probable	No Siniestros	% Siniestro
C14	Conducir desatento a las condiciones de transito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor).	5640	26,41
C23	No respetar las seniales reglamentarias de transito. (pare, ceda el paso, luz roja del semaforo, etc).	3861	18,08
C09	Conducir vehiculo superando los limites maximos de velocidad.	3049	14,28
C11	No mantener la distancia prudencial con respecto al vehiculo que le antecede.	1828	8,56
C06	Conduce bajo la influencia de alcohol, sustancias estupefacientes o psicotropicas y/o medicamentos.	1656	7,76
C12	No guardar la distancia lateral minima de seguridad entre vehiculos.	867	4,06
C19	Realizar cambio brusco o indebido de carril.	810	3,79
C18	Conducir en sentido contrario a la via normal de circulacion.	599	2,81
C25	No ceder el derecho de via o preferencia de paso a vehiculos.	541	2,53
C17	Bajarse o subirse de vehiculos en movimiento sin tomar las precauciones debidas.	504	2,36
	Otros	1997	9,36
	Total	**21352**	**100,00**

La causa probable dominante es *conducir desatento a las condiciones de tránsito*, con una consecuencia del 26.41% en los siniestros viales. Otras causas probables que inciden, en orden de influencia son: *no respectar las señales reglamentarias de tránsito, conducir superando los límites de velocidad, no mantener la distancia*

prudente con respecto al vehículo que le antecede, conducir bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas, no respectar la distancia lateral mínima de seguridad entre vehículos, realizar cambios bruscos o indebido de carril, no ceder el derecho de vía o preferencia de paso a los vehículos, bajarse o subirse de vehículos en movimiento sin tomar las precauciones debidas.

En la Tabla 3-4 se presenta las cusas probables de siniestros viales en relación con el tipo de siniestros. Aun cuando la causa probable de la influencia es la *conducir desatento a las condiciones de tránsito (celular, pantallas, comida, maquilla o cualquier otro distractor)* con 26.41 %, los tipos de siniestros que más influyen en esta casusa son: la perdida de pista con el 6.81%, el atropellamiento con 4.80% y los estrellamientos con el 4.63%.

Tabla 3-4 Causa Probable vs Tipo de Siniestro

Causa Probable en relación con el Tipo de Siniestro

Causa Probable	Arrollamientos	Atropellos	Caida de Pasajero	Choque Frontal	Choque Lateral	Choque Posterior	Colision	Estrellamientos	Otros	Perdida de Carril	Perdida de Pista	Rozamientos	Volcamientos	% TOTAL
C28					0,00									0,00
C27	0,07	0,77	0,00		0,01				0,00					0,85
C26	0,07	0,87			0,01				0,02			0,00		0,97
C25	0,00	0,08		0,10	2,16	0,03	0,01	0,01	0,07			0,06	0,00	2,52
C24	0,00	0,02			0,01	0,01		0,01	0,00		0,00	0,00		0,05
C23	0,01	0,35	0,03	0,23	16,11	0,19	0,06	0,36	0,58	0,04	0,04	0,07	0,03	18,10
C22		0,00		0,05	0,08	0,01	0,00	0,03	0,04	0,03		0,04	0,00	0,28
C21	0,01	0,05	0,01	0,04	0,05	0,01	0,01	0,07	0,02	0,10	0,07	0,01	0,02	0,47
C20		0,01		0,01	0,01	0,02		0,04	0,02	0,01	0,00	0,01	0,01	0,14
C19	0,01	0,08	0,03	0,23	2,65	0,07	0,01	0,17	0,06	0,15	0,10	0,21	0,01	3,78
C18		0,01		2,65	0,09			0,00	0,01	0,00	0,01	0,01		2,78
C17	0,01	0,01	2,33					0,00				0,00		2,35
C16	0,12	1,16	0,01		0,00	0,00		0,01	0,00	0,00	0,00	0,00		1,30
C15	0,00	0,01	0,07				0,00	0,01	0,00					0,09
C14	0,32	4,80	0,09	1,49	3,48	1,18	0,42	4,63	0,41	1,54	6,81	0,49	0,76	26,42
C12	0,00	0,01	0,00	0,02	0,72	0,07	0,03	0,07	0,07		0,00	3,05		4,04
C11	0,02	0,05		0,02	0,10	7,04	1,13	0,10	0,05	0,02	0,00	0,02	0,00	8,55
C10	0,00	0,10	0,01	0,08	0,27	0,13	0,03	0,52	0,12	0,22	0,03	0,01	0,07	1,59
C09	0,10	2,54	0,10	0,18	1,02	0,55	0,36	3,82	0,56	1,58	2,96	0,08	0,44	14,29
C08								0,01	0,02	0,00			0,02	0,05
C07	0,02	0,14	0,00		0,02			0,02	0,01		0,00	0,00		0,21
C06	0,03	0,40	0,01	0,64	2,09	0,98	0,37	1,76	0,16	0,31	0,60	0,33	0,06	7,74
C05	0,01	0,02		0,01	0,02	0,03	0,04	0,52	0,09	0,02	0,08	0,00	0,01	0,85
C04	0,01	0,01		0,01	0,01	0,02		0,05	0,04	0,05	0,03	0,00	0,01	0,24
C03		0,02		0,04	0,02	0,00	0,01	0,28	0,03	0,05	0,09	0,00	0,04	0,58
C02				0,03	0,01			0,03	0,01	0,06	0,04		0,02	0,20
C01	0,01	0,07	0,02	0,02	0,16	0,03	0,01	0,13	0,52	0,14	0,09	0,02	0,03	1,25
Total	**0,82**	**11,58**	**2,71**	**5,85**	**29,10**	**10,37**	**2,49**	**12,65**	**2,91**	**4,32**	**10,95**	**4,41**	**1,53**	**100,00**

Similarmente, la Tabla 3-4 muestra que la causa probable *C23 - No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc.)* y el tipo de siniestro *choque lateral* aportan en un 16.11% a la totalidad de siniestros reportados en Ecuador en el año 2021. También se presentan otros tipos de siniestros que influyen en las estadísticas como: el coche lateral con un 29.10%, los

estrellamientos con un 12.65%, los atropellamientos con 11.58%, el choque posterior con 10.37%, el choque frontal con 5.85% y la perdida de carril con 4.32%.

En la Tabla 3-5 se presentan las cusas probables en relación con los tipos de vehículos relacionados en los siniestros viales. Aun cuando la causa probable que más incide es la *conducir desatento a las condiciones de tránsito (celular, pantallas, comida, maquilla o cualquier otro distractor)* con 26.41 %, el tipo de vehículo automóvil tiene una influencia del 33.57 % en los accidentes de tránsito. También se presentan otros tipos de vehículos que influyen en las estadísticas de siniestros viales como: la motocicleta con 24.02 %, camiones con 6.84% y vehículo deportivo utilitario con el 5.13%, acabe señalar que los datos recolectados muestran que un 13.94 % de vehículos que inciden en la accidentabilidad no se han registrado. Igualmente, para la causa probable *C23 - No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc.)* y el tipo de vehículo automóvil tienen una influencia del 6.47%.

Tabla 3-5 Causa Probable vs Tipo de Vehículo

Causa Probable en relación con el Tipo de Vehiculo

Causa Probable	Automovil	Bicicleta	Bus	Camion	Camioneta	Emergencias	Especial	Furgoneta	Motocicleta	No Identificado	Scooter Electrico	Tricimoto	Vehiculo Deportivo Utilitario	% Total
C01	0,31	0,03	0,05	0,15	0,11	0,00	0,01	0,01	0,33	0,20		0,00	0,05	1,25
C02	0,03	0,01	0,01	0,02	0,03				0,08	0,02			0,00	0,20
C03	0,29		0,02	0,02	0,07			0,01	0,10	0,06			0,03	0,60
C04	0,08		0,02	0,04	0,06		0,00	0,00	0,01	0,01			0,01	0,23
C05	0,37	0,01	0,08	0,15	0,11		0,01	0,02	0,08				0,02	0,85
C06	4,45	0,04	0,10	0,33	1,00	0,01	0,02	0,09	1,15	0,12			0,43	7,74
C07	0,08	0,00	0,02		0,02			0,00	0,02	0,06				0,20
C08	0,00			0,03	0,02									0,05
C09	5,80	0,06	0,38	0,61	1,17	0,01	0,08	0,15	3,84	1,72	0,01	0,01	0,43	14,27
C10	0,81	0,01	0,06	0,11	0,16		0,01	0,04	0,24	0,07			0,09	1,60
C11	2,71	0,15	0,44	0,77	1,09	0,00	0,04	0,03	1,81	0,92		0,00	0,60	8,56
C12	1,21	0,10	0,23	0,37	0,30		0,04	0,01	1,14	0,49		0,00	0,15	4,04
C14	6,63	0,14	0,73	2,80	3,22	0,00	0,14	0,05	5,54	5,19		0,01	1,97	26,42
C15	0,01	0,01	0,07	0,00					0,01	0,00				0,10
C16	0,28	0,01	0,16	0,06	0,10		0,00	0,01	0,34	0,33		0,00	0,04	1,33
C17	0,04		0,32	0,03	0,01		0,00		0,24	1,70			0,00	2,34
C18	0,75	0,04	0,07	0,24	0,41	0,00	0,01	0,01	0,80	0,22			0,25	2,80
C19	1,33	0,05	0,15	0,15	0,28		0,01	0,04	1,43	0,23		0,01	0,10	3,78
C20	0,05		0,01	0,02	0,02		0,01		0,02	0,01			0,00	0,14
C21	0,07	0,02	0,02	0,05	0,04		0,01		0,15	0,10			0,01	0,47
C22	0,12	0,01	0,01	0,03	0,02				0,07	0,01			0,02	0,29
C23	6,47	0,21	0,59	0,80	1,74	0,02	0,05	0,11	5,25	1,94	0,00	0,05	0,85	18,08
C24	0,04		0,00		0,00				0,02	0,01				0,07
C25	1,07	0,09	0,06	0,03	0,28	0,01		0,01	0,86	0,05	0,00		0,07	2,53
C26	0,25	0,01	0,04	0,01	0,04		0,00	0,00	0,24	0,37			0,00	0,96
C27	0,32	0,01	0,09	0,02	0,04			0,02	0,25	0,11			0,01	0,87
C28	0,00													0,00
Total	**33,57**	**1,01**	**3,73**	**6,84**	**10,34**	**0,05**	**0,44**	**0,61**	**24,02**	**13,94**	**0,01**	**0,08**	**5,13**	**100,00**

De igual forma, la Tabla 3-6 muestra las causas probables con relación a las provincias, como se había mencionado en la descripción de datos las provincias con

mayores incidentes viales son: Guayas con 36.32%, Pichincha con 17.81% y Manabí con 9.15 %.

Tabla 3-6 Causa Probable vs Provincia

Causa Probable en relación con Provincias

Provincia	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26	C27	C28	Total
Azuay	0,04		0,02	0,00		0,42	0,01		0,11	0,04	0,38	0,12	1,31		0,04	0,03	0,24	0,03	0,00	0,01	0,03	0,75		0,15	0,05	0,06	0,00	3,84
Bolivar				0,00		0,02			0,00	0,02			0,32							0,00								0,36
Caniar	0,00					0,03			0,01				0,21		0,01		0,02											0,28
Carchi			0,00	0,00	0,01	0,06			0,12	0,01	0,00		0,09		0,01	0,00		0,00		0,01				0,00				0,31
Chimborazo	0,01	0,00	0,01	0,00	0,00	0,08	0,00		0,10	0,01	0,09	0,06	2,19		0,02		0,01	0,01	0,00	0,01		0,18	0,00	0,02	0,06	0,02		2,88
Cotopaxi	0,05			0,00		0,04	0,00		0,02	0,00			0,31		0,03		0,00			0,01		0,00						0,46
El Oro	0,01	0,01	0,02		0,01	0,27	0,00		0,28	0,00	0,39	0,13	0,95			0,02	0,16	0,09	0,01	0,04		1,22	0,01	0,01		0,02		3,65
Esmeraldas	0,02		0,02	0,01	0,02	0,04			0,21	0,01	0,06	0,11	0,30		0,00	0,00		0,05	0,01	0,01	0,00	0,13		0,02	0,01	0,04		1,07
Galapagos													0,00															0
Guayas	0,25	0,03	0,04		0,10	2,20	0,05	0,01	7,26		4,08	1,82	6,76	0,02	0,27	1,86	1,21	2,04	0,02	0,03	0,01	7,98	0,01	0,01	0,08	0,18		36,32
Imbabura	0,01	0,01	0,02	0,02	0,01	0,12	0,03		0,18	0,02	0,10	0,04	0,47	0,00	0,11	0,00				0,02	0,03	0,38	0,01	0,01	0,01			1,6
Loja	0,01		0,00	0,01	0,04	0,52	0,03		0,44	0,02	0,01		1,12		0,01		0,01		0,00	0,00	0,01	0,06	0,00	0,01		0,07		2,37
Los Rios	0,05	0,00	0,08	0,01	0,01	0,07			0,05	0,02	0,66	0,30	1,87		0,01	0,13	0,42	0,06	0,01	0,10	0,01	1,38		0,08	0,04			5,36
Manabi	0,56	0,01	0,05	0,07	0,04	0,12	0,00	0,01	0,34	0,00	1,01	0,66	3,13	0,00	0,07	0,13	0,24	0,57	0,01	0,01	0,02	1,05	0,02	0,86	0,14	0,03		9,15
Morona Santiago	0,01	0,00	0,02			0,04			0,02	0,03		0,00	0,41		0,07	0,00	0,01		0,00		0,00	0,01						0,62
Napo	0,01	0,02	0,00	0,00		0,02		0,00	0,05	0,01		0,00	0,07				0,00	0,01		0,00		0,00			0,00			0,19
Orellana						0,01			0,03	0,00			0,11					0,00	0,01	0,04		0,00						0,2
Pastaza	0,00					0,01			0,00	0,01			0,14				0,01			0,00				0,00				0,17
Pichincha	0,04	0,07	0,27	0,07	0,54	2,49	0,02	0,02	3,99	1,14	0,78	0,32	2,25	0,06	0,53	0,06	0,12	0,66	0,02	0,07	0,15	2,36	0,01	1,01	0,34	0,42		17,81
Santa Elena	0,03										0,28	0,10	1,01			0,06	0,13					0,82						2,43
Santo Domingo	0,12	0,03	0,03	0,02	0,03	0,25	0,01	0,00	0,43	0,19	0,69	0,26	0,90	0,01	0,01	0,04	0,20	0,21	0,02	0,10	0,01	0,93		0,31	0,24	0,02		5,06
Sucumbios						0,01				0,00			0,21															0,22
Tungurahua	0,04		0,00		0,03	0,95	0,04	0,00	0,60	0,03	0,02	0,14	2,13	0,01	0,11	0,01	0,00	0,05	0,01	0,00	0,01	0,81	0,00	0,02		0,00		5,01
Zamora Chinchipe		0,00	0,00	0,01		0,00	0,00		0,01	0,01			0,15		0,01		0,01	0,00										0,2
Total	**1,26**	**0,18**	**0,58**	**0,22**	**0,84**	**7,77**	**0,19**	**0,04**	**14,3**	**1,57**	**8,55**	**4,06**	**26,41**	**0,1**	**1,31**	**2,34**	**2,79**	**3,78**	**0,12**	**0,46**	**0,28**	**18,1**	**0,06**	**2,51**	**0,97**	**0,86**	**0**	**100**

En relación con los datos presentados en la Tabla 3-6 la provincia del Guayas tiene tres causas probables que inciden mayormente a la siniestralidad vial como: *C23 - No respetar las señales reglamentarias de tránsito. (pare, ceda el paso, luz roja del semáforo, etc.)* con 7.98%, *C09 - Conducir vehículo superando los límites máximos de velocidad* con 7.26%, *C14 - Conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor)* con 6.76 % *y C11 - No mantener la distancia prudencial con respecto al vehículo que le antecede* con 4.08%. Similarmente para la provincia de Pichincha la causa probable que inciden mayormente es *C09 - Conducir vehículo superando los límites máximos de velocidad* con 3.99%.

La exploración de datos de los siniestros viales ya puede inferir sobre las variables a considerar y los clústeres objetivos en la presente investigación aplicada

3.1.3. Preparación de los Datos

La preparación de los datos es una etapa fundamental para la presente investigación aplicada y con frecuencia la fase más recurrente y exigente en la minería de datos. Según IBM la fase de preparación de datos de la metodología CRIPS DM suele llevar el 50 % a 70 % del tiempo y esfuerzo de un proyecto de minería de datos. (IBM, 2021)

En consecuencia, se dedicará el tiempo y esfuerzo necesario para preparar e integrar los datos necesarios para la minería en lineamiento con las fases de compresión del tema de interés y compresión de los datos obtenidos de las fuentes secundarias, algunas de las tareas claves a realizar en esta fase son:

- Limpieza, eliminación o sustitución de datos.
- Agregación de datos.
- Integración o fusión de datos.
- Selección de muestras para análisis segregados.
- Clasificación de datos para el modelado.

3.1.3.1. Selección de Datos

En base a la descripción de los datos realizados es necesario considerar las siguientes anotaciones:

- Para el estudio de clústeres o agrupamiento mediante modelos no supervisados se conoce a priori que el conjunto de datos no dispone de datos etiquetados, no obstante, se buscará agrupar los objetos según su similaridad en base a sus características y propiedades de distancia.
- Se considera únicamente los sinestros viales registrados en el año 2021 por las entidades públicas del Ecuador y de relación con la seguridad vial en Ecuador como el Ministerio de Transporte y Obras Públicas (MTOP), Comisión de Tránsito del Ecuador (CTE), Gobiernos Autónomos Descentralizados Municipales (GADMs) y la Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Vial (ANT).
- Según la descripción y exploración de datos realizado al conjunto de datos, se decide acotar este análisis a las provincias que tienen mayor porcentaje de incidencia en los siniestros viales y sus respectivas características como: causa probable, tipo de siniestros, tipo de vehículo, sexo, edad, condición (fallecido, lesionado), cantón y parroquia del accidente registrado, se

muestra en la Tabla 3-7 las tres provincias con mayor incidencia en siniestros viales.

Tabla 3-7 Provincias con más incidencia en Siniestros Viales en Ecuador en el año 2021

Provincia	Incidentes	Procentaje
Guayas	7758,00	36,33
Pichincha	3807,00	17,83
Manabi	1957,00	9,17
Los rios	1145,00	5,36
Santo Domingo de los Tsachilas	1081,00	5,06
Tungurahua	1076,00	5,04
Azuay	828,00	3,88
El oro	783,00	3,67
Chimborazo	622,00	2,91
Santa Elena	517,00	2,42
Loja	512,00	2,40

3.1.3.2. Limpieza de Datos

Para la conjunto datos de siniestros viales registrados se han efectuado algunas técnicas de limpieza con la finalidad de tener calidad en la información.

1. Se excluyen todos los registros que no contienen información sobre las características seleccionadas para el presente caso de estudio como: *causa probable, tipo de siniestros, tipo de vehículo, sexo, edad, condición (fallecido, lesionado), cantón y parroquia del accidente registrado.*
2. Se toman la *longitud, latitud* en base a la *provincia* y la *fecha* del incidente.
3. Se reemplazan los caracteres especiales, caracteres con acentos para evitar problemas de codificación en las herramientas utilizadas.
4. Se crean nuevas variables a partir de la fecha del registro del siniestro vial como: *periodo* (segmentado por rango de horas), *día* (nombre del día y su presentación número) y *mes* (nombre del mes y su representación numérica).
5. Se crean nuevas variables dummy que es una variable binaria que indica si una variable categórica independiente adopta un valor específico en las variables *Tipo de Siniestro* y *Tipo de Vehículo*.

6. Se toma el esquema UTF-8 como codificación de facto para el conjunto de datos.

Una vez ejecutados las técnicas de limpieza, se obtiene un conjunto de datos de 21.352 registros.

3.1.3.3. Integración de Datos

A partir de la comprensión del tema de interés, se integró datos de varias fuentes secundarias asociadas a entidades públicas como se muestra en la Tabla 3-8.

Tabla 3-8 Entidades relacionadas al Control Vial en Ecuador

Código	Entidad de Control
PNE	PNE.- Policía Nacional de Ecuador.
DNCTSV	Dirección Nacional de Control de Tránsito y Seguridad Vial (Policía Nacional).
CTE	Comisión de Tránsito del Ecuador.
ATM	Autoridad de Tránsito Municipal de Guayaquil.
AMT	Agencia Metropolitana de Tránsito de Quito.
MEP	Mancomunidad del Norte Empresa Pública.
MCU	Municipio de Cuenca EMOV EP.
MAM	Municipio de Ambato.
MLO	Municipio de Loja UCOT.
MSD	Municipio de Santo Domingo.
MMA	Municipio de Manta.
MPO	Municipio de Portoviejo.
MMA	Municipio de Babahoyo.
MMC	Municipio de Machala.
MRI	Municipio de Riobamba.
MDI	Ministerio del Interior.
OISEVI	Observatorio Iberoamericano de Seguridad Vial .
OMS	Organización Mundial de la Salud.
OPS	Organización Panamericana de la Salud.
LOTTTSV	Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial.
RLOTTTSV	Reglamento a Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial.
INEC	Instituto Nacional de Estadísticas y Censos.
INEN	Servicio Ecuatoriano de Normalización.
MTOP	Ministerio de Transportes y Obras Públicas.

Cabe señalar que el reporte nacional de siniestralidad vial que su mayoría es recolecta e integrada en los partes policiales, los mismos que son diseñados y aprobados por las entidades de control, bajo los parámetros técnicos establecidos por la Agencia Nacional de Tránsito (ANT), en tan sentido el agente de tránsito levanta un parte policial que certifica la ocurrencia de un siniestro de tránsito y posteriormente se realizan dos procesos:

1. Ingreso y validación de datos al Sistema Nacional de Estadísticas de Tránsito SINET, plataforma interna de la ANT.
2. Publicación y libre acceso a los datos en el Sistema Nacional de Estadísticas de Tránsito SINET.

Finalmente, la información ingresada es validada por parámetros de consistencia en el Sistema Nacional de Estadísticas de Tránsito SINET cuyo fin es busca y detectar posibles errores de consistencia u omisión de información.

3.1.3.4. Formato de Datos

Como paso final en la preparación de datos y antes de la generación del modelo, se ha validado que cada uno de los tributos tenga el formado adecuado para el procesamiento de la información, se presenta en la Tabla 3-9 el formato determinado para las variables del conjunto de datos.

Tabla 3-9 Variables del Conjunto de Datos

Variable	Tipo Dato		Variable	Tipo Dato
ANIO	int64		FERIADO	object
SINIESTROS	object		CODIGO_CAUSA	object
LESIONADOS	int64		CAUSA_PROBABLE	object
FALLECIDOS	int64		TIPO_DE_SINIESTRO	object
ENTE_DE_CONTROL	object		TIPO_DE_VEHICULO_1	object
LATITUD_Y	float64		SERVICIO_1	object
LONGITUD_X	float64		AUTOMOVIL	int64
DPA_1	int64		BICICLETA	int64
PROVINCIA	object		BUS	int64
DPA_2	int64		CAMION	int64
CANTON	object		CAMIONETA	int64
DPA_3	int64		EMERGENCIAS	int64
PARROQUIA	object		ESPECIAL	int64
DIRECCION	object		FURGONETA	int64
ZONA_PLANIFICACION	object			int64
ZONA	object		NO_IDENTIFICADO	int64
ID_DE_LA_VIA	object		SCOOTER_ELECTRICO	int64
NOMBRE_DE_LA_VIA	object		TRICIMOTO	int64
UBICACION_DE_LA_VIA	object		VEHICULO_DEPORTIVO_UTILITARIO	int64
JERARQUIA_DE_LA_VIA	object		SUMA_DE_VEHICULOS	int64
FECHA	datetime64		TIPO_ID_1	object
HORA	timestamp		EDAD_1	int64
PERIODO_1	object		SEXO_1	object
PERIODO_2	int64		CONDICION_1	object
DIA_1	object		PARTICIPANTE_1	object
DIA_2	int64		CASCO_1	object
MES_1	object		CINTURON_1	object
MES_2	int64			

3.1.4. Modelado

Con el objetivo de proveer el entendimiento abordado en el tema de interés y apalancado con la metodología plateada en la presente tesis aplicada, se propone realizar el estudio con tres modelos no supervisados, para encontrar los clústeres y características de los sinestros viales que más inciden en la accidentabilidad del Ecuador en el año 2021. Por lo tanto, en este capítulo se profundizará la aplicación de los modelos seleccionados en el conjunto de datos, las características seleccionadas la explicación sobre las técnicas de evaluación de los modelos, la construcción y características de los modelos y la presentación de los resultados.

3.1.4.1. Selección de técnicas de modelado

Fundamentado en la metodología planteada, para la presente tesis aplicada se realiza una comparación entre los tres modelos no supervisados. El primer modelo *K-means* cuyo objetivo es la partición del conjunto de n datos en k grupos, basándose en sus características. Por lo general el agrupamiento se realiza minimizando la suma de distancias cuadrática entre cada observación y el centroide del clúster. Este primer modelo servirá como punto de referencia y comparación de los modelos restantes. Por otro lado, en cuanto a modelos de aprendizaje no supervisados se debe considerar que los datos no están etiquetados y estos se clasifican partiendo de su estructura interna como sus propiedades o características. Por otro lado, también se aborda el tema de interés utilizando el modelo *Agrupación Jerárquica* cuya finalidad es agrupar los datos en un árbol de clústeres donde cada observación se considera como un clúster independiente y por último el modelo *DBSCAN* que identifica los clústeres basándose en la densidad de aplicaciones.

3.1.4.2. Generación de modelos

A continuación, se especifican los parámetros y hiperparámetros que son utilizados para ajustar y modelar cada uno los algoritmos no supervisados que se utilizarán en la presente investigación aplicada. Como se ha planteado en la preparación de datos el conjunto de datos contiene información de número de siniestros ocurrido en Ecuador en el año 2021 y las principales características recogidas por las entidades de control.

Como modelo base se han seleccionado todas las variables que influyen en los siniestros que se muestran en la Tabla 3-10.

Tabla 3-10 Variables del Modelo

Variable	Tipo Dato
CANTON	object
ZONA	object
PERIODO	object
DIA	object
MES	object
CODIGO_CAUSA	object
TIPO_DE_VEHICULO	object
SEXO	object
FERIADO	object

- Algoritmo K-means: establece agrupaciones por particiones y se debe determinar preliminarmente el número de clústeres k, cada clúster tiene asignado un centroide (centro geométrico), se asignan los puntos a cada clúster utilizando una métrica de distancia, el algoritmo iterativamente va asignando los puntos al clúster, este proceso se realiza en las provincias seleccionadas para el estudio que se muestra en la Tabla 3-7.
  i. Inicialización
     - Establecer k centroides en base al método de elbow.
     - Formar k grupos, asignando cada observación al centroide más cercano.
  ii. Proceso Iterativo
     - Calcular las distancias de todos los puntos a los k centroides.
  iii. Complejidad
     - n cantidad de observaciones.
     - k número de clústeres.
     - i número de iteraciones.
     - d número de atributos.
- Algoritmo Agrupación Jerárquica: separa el clúster basado en la similaridad, ejecutando iterativamente el proceso aglomerativo o divisivo,

minimizando la distancia o maximizando alguna característica de similitud. Para la presente investigación aplicada se utilizará el proceso *aglomerativo* que considera a las observaciones un clúster particular y único al principio, este proceso se realiza en las provincias seleccionadas para el estudio que se muestra en la Tabla 3-7

i. Inicialización
   - Cálculo de la matriz de proximidad
   - La similitud de la observación esta dado por la altura del nodo común más cercano.

ii. Proceso Iterativo
   - Determinar el par de clústeres más similares o diferentes en términos de distancia, haciendo uso de la matriz de proximidad, formando grupos cada vez más grandes y heterogéneos.
   - Actualización de la matriz de proximidad.

iii. Complejidad
   - Determinar la medida para el cálculo de las distancias existente entre los elementos a fusionar, existen 7 distintos tipos de medios, para la presente investigación aplicada se utilizó el método *ward*
   - *Ward* es la técnica más utilizada debido a que maximiza la homogeneidad de las agrupaciones.

- Algoritmo DBSCAN: basado en la densidad, comúnmente utilizado cuando existen datos irregulares con ruido, este proceso se realiza en las provincias seleccionadas para el estudio que se muestra en la Tabla 3-7.

i. Inicialización
   - Establecer parámetro eps.
   - Establecer parámetros min_sample.

ii. Proceso Iterativo
   - Calcular la matriz de distancias entre las observaciones.

- Clasificar cada observación en base a punto central, punto frontera y ruido.

iii. Complejidad

- eps radio máximo de vecindad, donde dos observaciones se consideran vecinos si la distancia entre ambos es inferior a eps
- min_sample número mínimo de observaciones para formar un clúster, considerando eps.

3.1.4.3. Evaluación del modelo

En cuanto a las técnicas de evaluación se ha normalizado a priori las características seleccionadas con la finalidad de que una variable no domine sobre otras. Asimismo, se consideran las siguientes técnicas de evolución de mejor ajuste a los modelos:

Validación Externa: establece la calidad del agrupamiento conociendo la información del conjunto de datos, es decir se conocen las etiquetas del conjunto de datos. Este tipo de métricas no serán aplicadas al conjunto de datos en la presente tesis de investigación aplicada.

Validación Interna: estables métricas de calidad del agrupamiento sin la necesidad de conocer a priori las etiquetas de conjunto de datos, de forma que se pueda seleccionar el mejor modelo, así como el numero óptimo de clústeres.

- *Cohesión*: Minimizar la distancia intra-clúster, es decir cada observación en el clúster debe ser los más cercano a los miembros del clúster.
- *Separación*: Maximizar la distancia inter-clúster, es decir los clústeres deben estar ampliamente separados entre ellos. Hay varios enfoques para la validación inter-clúster: distancia entre la observación más cercana, distancia entre las observaciones más lejanas o la distancia entre centroides.
- *SSW - Sum of Squared Within*: métrica para evaluar la cohesión de los clústeres determinados por el algoritmo.
- *SSB - Sum of Squared Between*: métrica para evaluar la separación de los clústeres determinado por el algoritmo.

Los índices o métricas basadas en la suma de cuadrados se caracterizan por cuantificar la dispersión de las observaciones a nivel inter-clúster e intra-clúster como:

I. Índice Davies Bouldin

   El índice de Davies-Bouldin mide el tamaño de los clústeres en función de la distancia media entre ellos. Un índice bajo significa que los clústeres están mejor definidos.

II. Índice Calinski - Harabasz

   El índice Calinski - Harabasz mide la dispersión inter-clústeres frente a la dispersión intra-clústeres. Un índice alto significa que los clústeres están mejor definidos.

III. Coeficiente Silhouette

   El coeficiente de silueta mide la distancia inter-clústeres resultantes. Un coeficiente más alto significa que el clúster está lejos del clúster más cercano y que el clúster está mejor definido.

En la Tabla 3-11 se muestra el índice y la ecuación de cálculo de las métricas.

Tabla 3-11 Índices para Evaluación de Calidad de Clústeres

Índice	Formula
* Davies Bouldin	$\frac{1}{k}\sum_{i=1,i\neq j}^{k} \max(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$
** Calinski y Harabasz	$\frac{SSB/(k-1)}{SSW/(n-k)}$

* Donde k es el número de clústeres, σi es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, σj es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y $d(ci\ ,\ cj)$ es la distancia entre los centroides de los 2 clústeres .

** Donde k el numero de clústeres, N el número de datos y d la dimension de los datos

Capitulo 4

4. Resultados

En este capítulo se presentan los resultados de los modelos de aprendizaje automático no supervisados aplicados en el tema de interés. *Cabe mencionar que las métricas que se utilizaron para evaluar los modelos son la cohesión y separación, de igual manera considerar que se acotó los resultados a las tres provincias con mayor incidencia en siniestros viales como se muestra en la Tabla 3-7.*

4.1. Análisis y Validación de Resultados

Cada algoritmo utilizado para responder al tema de interés tiene ventajas, desventajas y limitaciones debidas a las condiciones de los datos o a las especificaciones de sus parámetros como se abordó en el Marco Teórico. Cabe resaltar que teóricamente es complejo suponer que algoritmo es mejor que otro o cual es el mejor para inferir sobre el conjunto datos sin etiquetas, asimismo la aplicación de técnicas no supervisadas sobre datos no etiquetados no tienen una única solución, sin embargo los modelos no supervisados aplicados en la presente trabajo de investigación aplicada permite caracterizarlos mediante agrupaciones y obtener la información u conocimiento sobre cada clúster, a continuación, se expone los resultados obtenidos.

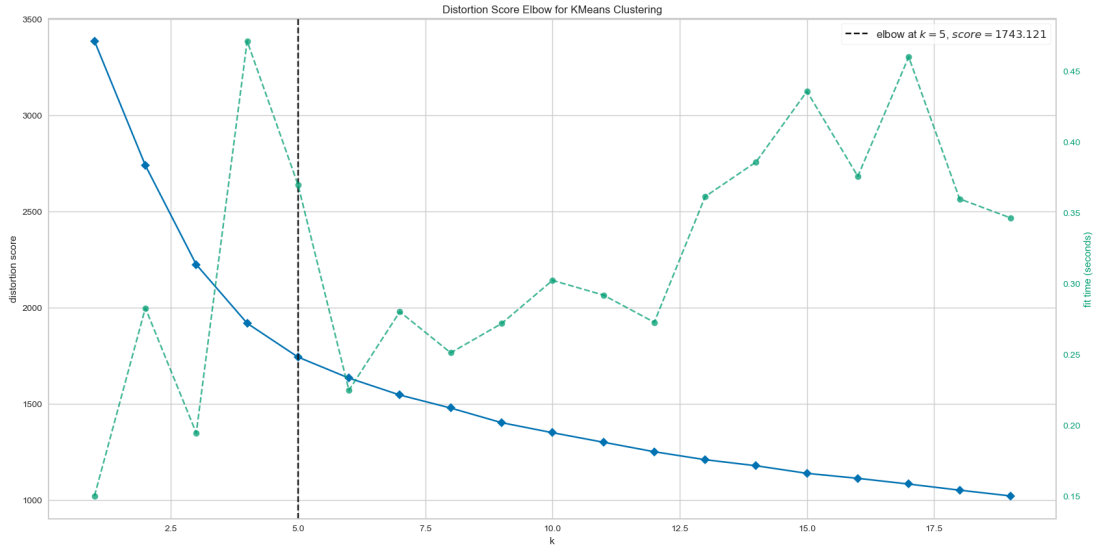
4.1.1. K-means

Utilizando el algoritmo de K-means se obtuvo los siguientes resultados, es necesario destacar que el modelo K-means minimiza la suma de distancias entre las observaciones y su respectivo centroide de clúster.

4.1.1.1. Guayas

La provincia de Guayas representa el 36.33 % de siniestro viales ocurridos en Ecuador para el año 2021, donde se han registrado 7.758 siniestros. Aplicando el método de Elbow sobre el subconjunto de datos se obtiene que el número óptimo de clústeres es $k = 5$, en la Figura 4-1 se muestra los resultados.

Figura 4-1 Método de Elbow para el valor óptimo de k en K-means, Guayas

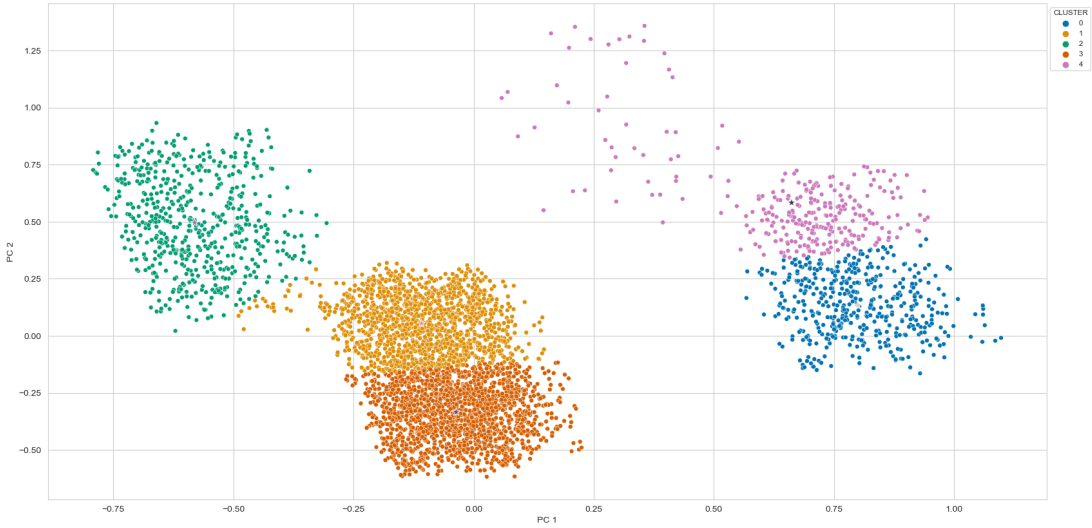


Fuente (Maza. E, 2022)

Del mismo modo, para obtener el número óptimo de clústeres k se utilizó la técnica de *distorsión* que se calcula como la media de las distancias al cuadrado de los centroides de los respectivos clústeres empleando la métrica de distancia euclidiana.

Por lo tanto, $k = 5$ óptimo de clústeres ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente Figura 4-2 muestra la visualización de clústeres para la provincia de Guayas.

Figura 4-2 K-means clústeres, Guayas



Fuente (Maza. E, 2022)

La siguiente Tabla 4-1, cataloga las principales características para los clústeres obtenidos con el modelo K-means en la provincia de Guayas.

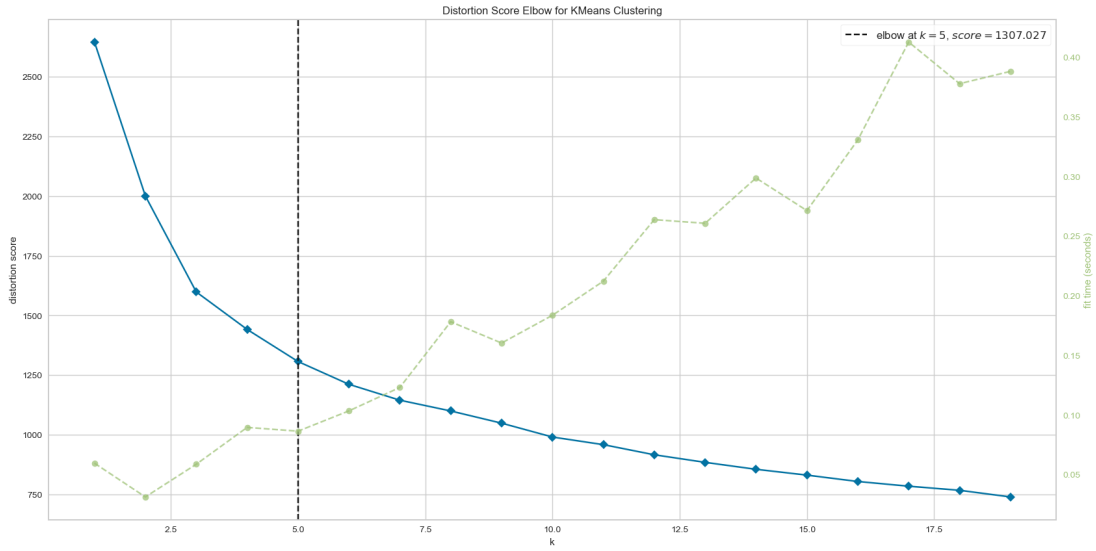
Tabla 4-1 Principales características por clúster en la provincia de Guayas, K-means

	CLUSTER				
CARACTERÍSTICAS	**0**	**1**	**2**	**3**	**4**
ZONA	Rural	Urbana	Urbana	Rural	Urbana
SEXO	Hombre	Hombre Mujer	Hombre	Hombre Mujer	Mujer
FERIADO	Si No	Si No	Si No	Si No	Si No
TIPO VEHÍCULO	Motocicleta Automovil Camion Camioneta Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Bus Camion	Automovil Moticicleta Bus Camioneta	Motocicleta Automovil Camion Camioneta Vehiculo Deportivo Utilitario	Automovil Motocicleta Bus Camioneta Vehiculo Deportivo Utilitario
TIPO CHOQUE	Perdida De Pista Choque Lateral Choque Posterior Estrellamientos Choque Frontal	Choque Lateral Atropellos Choque Posterior Perdida De Pista Estrellamientos	Choque Lateral Perdida De Pista Choque Posterior Estrellamientos Atropellos	Choque Lateral Perdida De Pista Choque Posterior Estrellamientos Choque Frontal	Choque Lateral Atropellos Perdida De Pista Choque Posterior Rozamientos
CAUSA PROBABLE	C14 C11 C23 C18 C12	C23 C09 C19 C11 C12	C09 C23 C06 C11 C19	C14 C23 C11 C18 C12	C09 C23 C11 C19 C06
MES	Enero Julio Junio Octubre Septiembre	Marzo Junio Enero Mayo Abril	Agosto Diciembre Octubre Enero Noviembre	Enero Marzo Junio Diciembre Abril	Agosto Junio Diciembre Febrero Noviembre
DIA	Sabado Domingo Viernes Jueves Miercoles	Lunes Miercoles Martes Jueves Domingo	Sabado Domingo Viernes Jueves Miercoles	Lunes Martes Miercoles Jueves Domingo	Jueves Viernes Martes Miercoles Lunes
PERIODO	De 06H00 A 06H59 De 22H00 A 22H59 De 13H00 A 13H59 De 16H00 A 16H59 De 19H00 A 19H59	De 15H00 A 15H59 De 16H00 A 16H59 De 14H00 A 14H59 De 17H00 A 17H59 De 07H00 A 07H59	De 07H00 A 07H59 De 08H00 A 08H59 De 18H00 A 18H59 De 19H00 A 19H59 De 15H00 A 15H59	De 18H00 A 18H59 De 22H00 A 22H59 De 20H00 A 20H59 De 17H00 A 17H59 De 19H00 A 19H59	De 16H00 A 16H59 De 20H00 A 20H59 De 18H00 A 18H59 De 14H00 A 14H59 De 17H00 A 17H59
CANTON	Daule Guayaquil San Jacinto De Yaguachi Naranjal Duran	Guayaquil Milagro Duran Samborondon Daule	Guayaquil Milagro Duran Daule Samborondon	Daule Guayaquil Naranjal Duran San Jacinto De Yaguachi	Guayaquil Milagro Samborondon Duran Daule

4.1.1.2. Pichincha

La provincia de Pichincha representa el 17.83 % de siniestro viales ocurridos en Ecuador para el año 2021, donde se han registrado 3.807 siniestros. Aplicando el método de Elbow sobre el subconjunto de datos se obtiene que el número óptimo de clústeres es $k = 5$, en el Figura 4-3 se muestra los resultados.

Figura 4-3 Método de Elbow para el valor óptimo de k en K-means, Pichincha

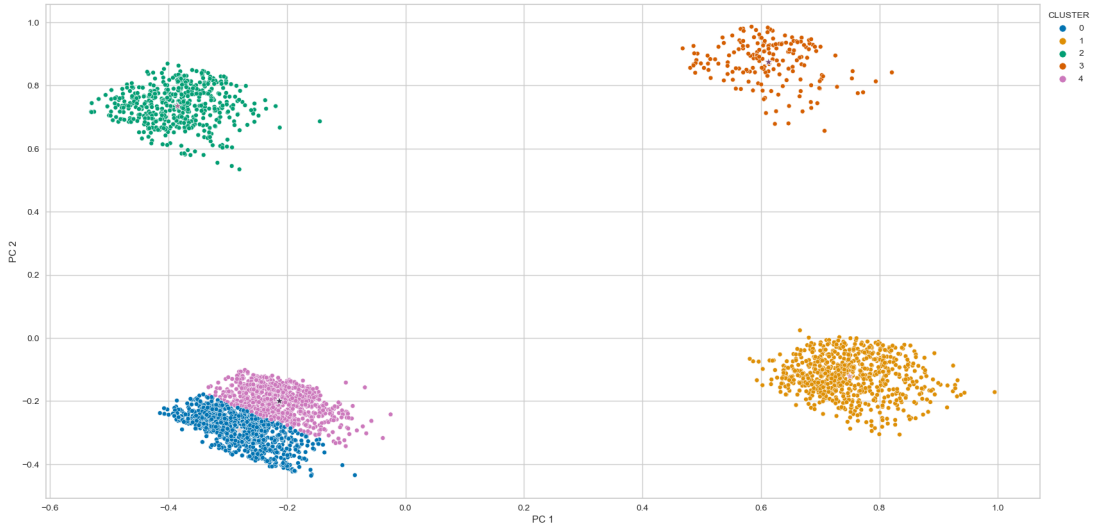


Fuente (Maza. E, 2022)

Del mismo modo, para obtener el número óptimo clústeres k se utilizó la técnica de *distorsión* que se calcula como la media de las distancias al cuadrado de los centroides de los respectivos clústeres empleando la métrica de distancia euclidiana.

Por lo tanto, $k = 5$ óptimo de clústeres ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente Figura 4-4 muestra la visualización de clústeres para la provincia de Pichincha.

Figura 4-4 K-means clústeres, Pichincha



Fuente (Maza. E, 2022)

La Tabla 4-2, cataloga las principales características para los clústeres obtenidos con el modelo K-means en la provincia de Pichincha.

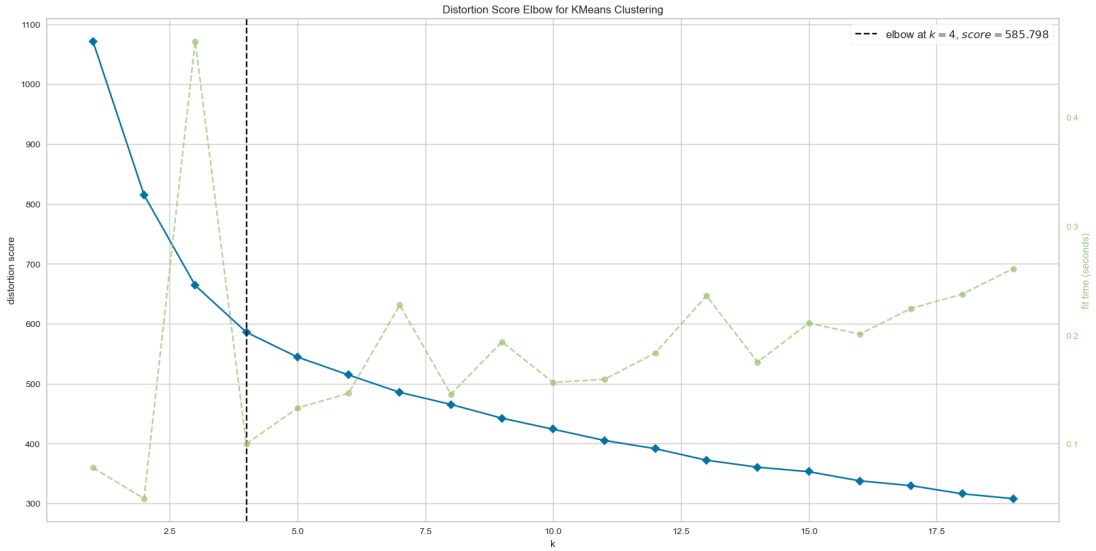
Tabla 4-2 Principales características por clúster en la provincia de Pichincha, K-means

	CLUSTER				
CARACTERÍSTICAS	**0**	**1**	**2**	**3**	**4**
ZONA	Urbana	Rural	Urbana	Rural	Urbana
SEXO	Hombre	Hombre	Mujer	Mujer	Hombre
FERIADO	No Si	No Si	No Si	No Si	No Si
TIPO VEHÍCULO	Automovil Motocicleta Camioneta Bus Camion	Automovil Motocicleta Camion Camioneta Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Bus Camion	Automovil Motocicleta Camioneta Vehiculo Deportivo Utilitario Bus	Automovil Motocicleta Camioneta Bus Camion
TIPO CHOQUE	Choque Lateral Estrellamientos Choque Posterior Atropellos Otros	Choque Lateral Estrellamientos Choque Posterior Perdida De Carril Choque Frontal	Choque Lateral Estrellamientos Atropellos Choque Posterior Colision	Choque Lateral Perdida De Carril Estrellamientos Atropellos Choque Frontal	Choque Lateral Estrellamientos Choque Posterior Atropellos Colision
CAUSA PROBABLE	C06 C09 C23 C10 C25	C14 C06 C09 C23 C10	C23 C09 C06 C10 C25	C14 C06 C09 C23 C10	C06 C23 C09 C11 C25
MES	Diciembre Octubre Noviembre Septiembre Agosto	Agosto Septiembre Marzo Enero Julio	Septiembre Julio Agosto Junio Diciembre	Enero Marzo Agosto Febrero Junio	Enero Marzo Febrero Mayo Abril
DIA	Sabado Domingo Viernes Miercoles Jueves	Domingo Sabado Viernes Jueves Miercoles	Viernes Sabado Domingo Jueves Miercoles	Sabado Domingo Viernes Jueves Lunes	Sabado Viernes Domingo Miercoles Martes
PERIODO	De 14H00 A 14H59 De 18H00 A 18H59 De 15H00 A 15H59 De 19H00 A 19H59 De 21H00 A 21H59	De 19H00 A 19H59 De 16H00 A 16H59 De 20H00 A 20H59 De 17H00 A 17H59 De 06H00 A 06H59	De 08H00 A 08H59 De 15H00 A 15H59 De 21H00 A 21H59 De 18H00 A 18H59 De 13H00 A 13H59	De 12H00 A 12H59 De 15H00 A 15H59 De 18H00 A 18H59 De 19H00 A 19H59 De 21H00 A 21H59	De 12H00 A 12H59 De 17H00 A 17H59 De 16H00 A 16H59 De 14H00 A 14H59 De 19H00 A 19H59
CANTON	Quito Ruminiahui Mejia Cayambe Pedro Moncayo	Quito Mejia Cayambe Pedro Moncayo Ruminiahui	Quito Ruminiahui Cayambe	Quito Mejia Pedro Moncayo Cayambe Ruminiahui	Quito Ruminiahui Cayambe Mejia Pedro Moncayo

4.1.1.3. Manabí

La provincia de Manabí representa el 9.17 % de siniestro viales ocurridos en Ecuador para el año 2021, donde se han registrado 1.957 siniestros. Aplicando el método de Elbow sobre el subconjunto de datos se obtiene que el número óptimo de clústeres es $k = 4$, en el Figura 4-3 se muestra los resultados.

Figura 4-5 Método de Elbow para el valor óptimo de k en K-means, Manabí

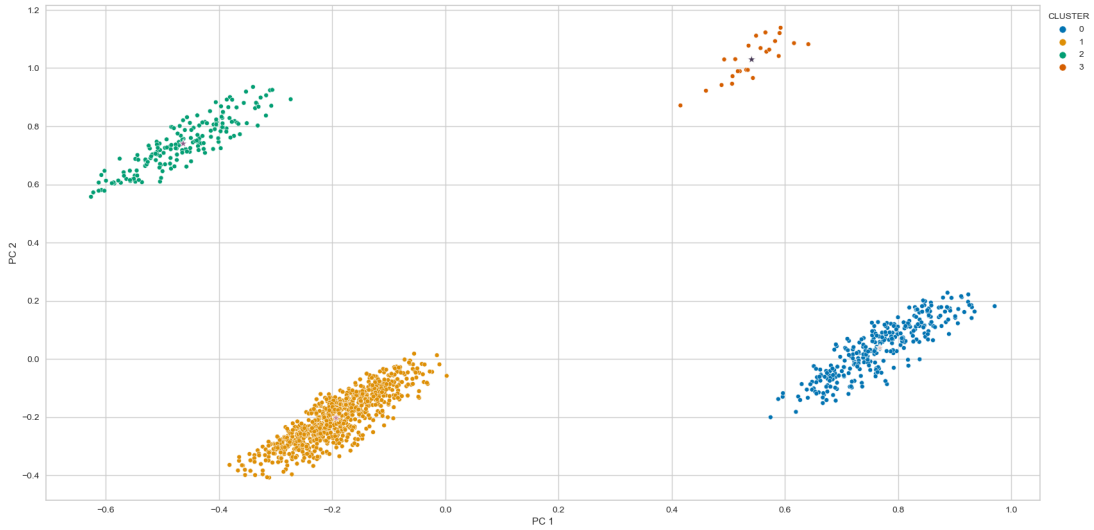


Fuente (Maza. E, 2022)

Del mismo modo, para obtener el número óptimo clústeres k se utilizó la técnica de *distorsión* que se calcula como la media de las distancias al cuadrado de los centroides de los respectivos clústeres empleando la métrica de distancia euclidiana.

Por lo tanto, $k = 4$ óptimo de clústeres ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente Figura 4-6 muestra la visualización de clústeres para la provincia de Manabí.

Figura 4-6 K-means clústeres, Manabí



Fuente (Maza. E, 2022)

La Tabla 4-3, cataloga las principales características para los clústeres obtenidos con el modelo K-means en la provincia de Pichincha.

Tabla 4-3 Principales características por clúster en la provincia de Manabí, K-means

	CLUSTER			
CARACTERÍSTICAS	**0**	**1**	**2**	**3**
ZONA	Rural	Urbana	Urbana	Rural
SEXO	Hombre	Hombre	Mujer	Mujer
FERIADO	No Si	No Si	No Si	No Si
TIPO VEHÍCULO	Motocicleta Automovil Camioneta Camion Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Camion Bicicleta	Automovil Motocicleta Camioneta Vehiculo Deportivo Utilitario Bicicleta	Automovil Motocicleta Camioneta Vehiculo Deportivo Utilitario
TIPO CHOQUE	Choque Lateral Perdida De Pista Choque Posterior Estrellamientos Choque Frontal	Choque Lateral Choque Posterior Choque Frontal Rozamientos Estrellamientos	Choque Lateral Choque Posterior Estrellamientos Rozamientos Atropellos	Choque Lateral Perdida De Pista Choque Frontal Estrellamientos Perdida De Carril
CAUSA PROBABLE	C14 C23 C11 C12 C18	C14 C25 C19 C11 C12	C14 C25 C11 C12 C19	C14 C23 C04 C09 C01
MES	Enero Marzo Junio Octubre Agosto	Diciembre Agosto Noviembre Septiembre Abril	Junio Septiembre Julio Agosto Diciembre	Enero Abril Febrero Mayo Agosto
DIA	Domingo Sabado Jueves Viernes Martes	Sabado Viernes Domingo Jueves Miercoles	Domingo Viernes Jueves Martes Lunes	Domingo Sabado Viernes Lunes Miercoles
PERIODO	De 17H00 A 17H59 De 20H00 A 20H59 De 19H00 A 19H59 De 12H00 A 12H59 De 18H00 A 18H59	De 17H00 A 17H59 De 13H00 A 13H59 De 19H00 A 19H59 De 15H00 A 15H59 De 20H00 A 20H59	De 14H00 A 14H59 De 11H00 A 11H59 De 16H00 A 16H59 De 19H00 A 19H59 De 17H00 A 17H59	De 17H00 A 17H59 De 18H00 A 18H59 De 09H00 A 09H59 De 15H00 A 15H59 De 19H00 A 19H59
CANTON	Portoviejo Montecristi Rocafuerte Chone Manta	Manta Portoviejo El Carmen Jipijapa Pedernales	Manta Portoviejo El Carmen Jaramijo Jipijapa	Jipijapa Montecristi Pedernales Portoviejo Santa Ana

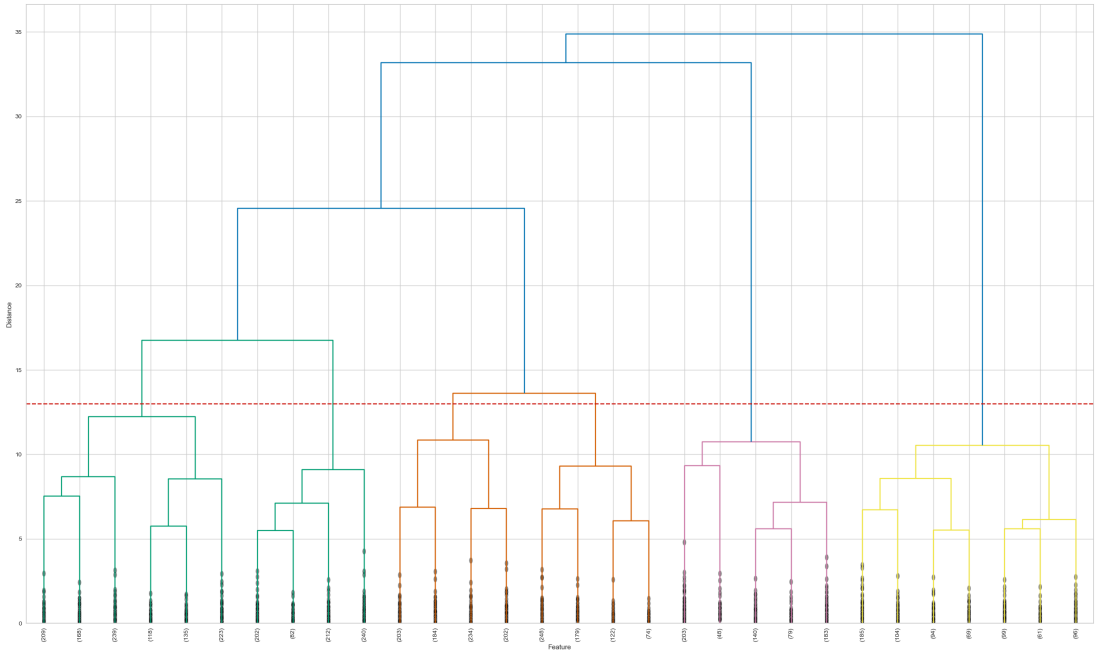
4.1.2. Agrupación Jerárquica.

Utilizando el algoritmo de agrupación jerárquica mediante la estrategia aglomerativa, se obtuvo los siguientes resultados, es necesario destacar que el modelo de agrupación jerárquica agrupa los datos con características similares.

4.1.2.1. Guayas

Utilizando la estrategia aglomerativa del modelo de Agrupación Jerárquico, sobre el conjunto de datos se obtuvo que el número óptimo de conglomerados para Guayas es $k = 6$, en la Figura 4-7 se muestra el dendrograma obtenido.

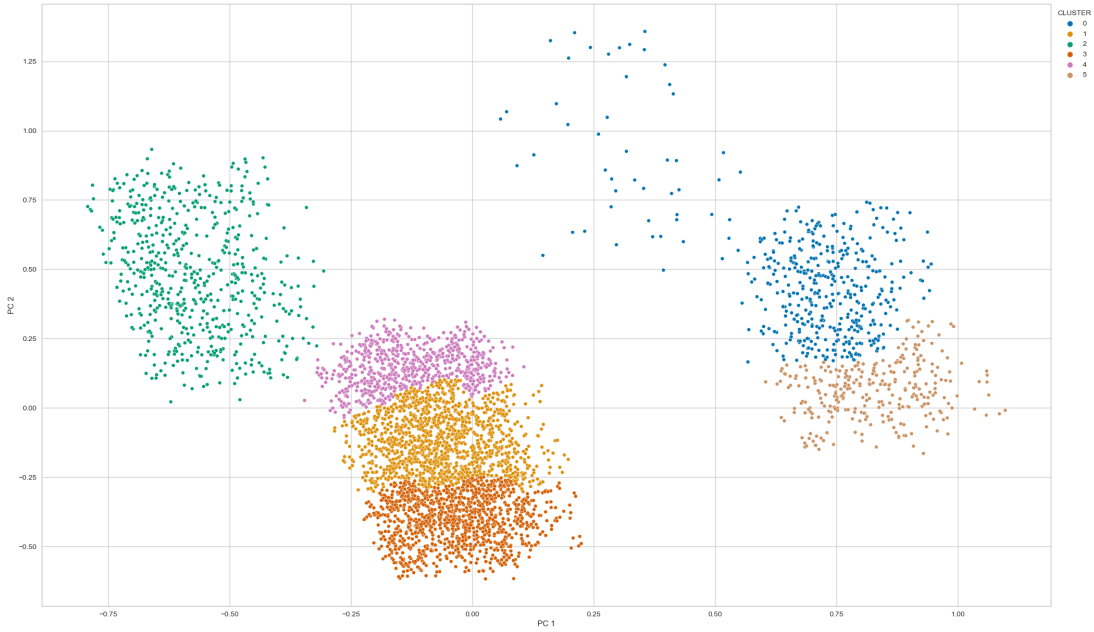
Figura 4-7 Dendrograma para el valor óptimo de k en Agrupación Jerárquica, Guayas



Fuente (Maza. E, 2022)

Por lo tanto, $k = 6$ óptimo de clústeres ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente Figura 4-8 muestra la visualización de clústeres para la provincia de Guayas.

Figura 4-8 Agrupación Jerárquica clústeres, Guayas



Fuente (Maza. E, 2022)

La Tabla 4-4, cataloga las principales características para los clústeres obtenidos con el modelo Agrupación Jerárquica en la provincia de Guayas.

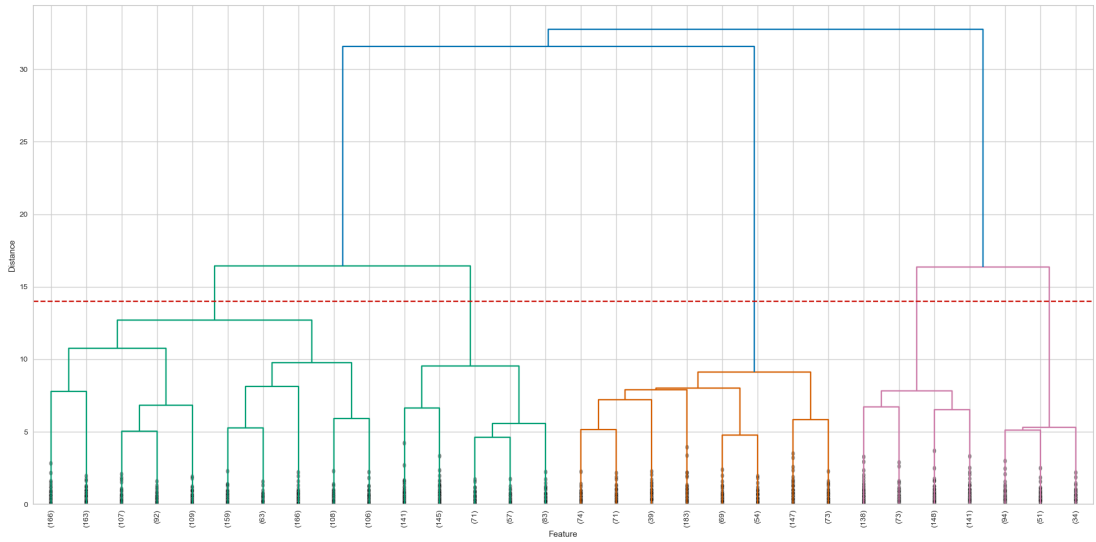
Tabla 4-4 Principales características por clúster en la provincia de Guayas, Agrupación Jerárquica

	CLUSTER					
CARACTERÍSTICAS	**0**	**1**	**2**	**3**	**4**	**5**
ZONA	Rural	Urbana	Urbana	Urbana	Urbana	Rural
SEXO	Hombre Mujer	Hombre	Mujer	Hombre	Hombre Mujer	Hombre
FERIADO	No Si	No Si	No Si	No Si	No Si	No Si
TIPO VEHÍCULO	Automovil Motocicleta Camion Camioneta Vehiculo Deportivo Utilitario	Motocicleta Automovil Camioneta Vehiculo Deportivo Utilitario Bus	Automovil Motocicleta Bus Camioneta Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Camion Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Bus Camion	Motocicleta Automovil Vehiculo Deportivo Utilitario Camion Camioneta
TIPO CHOQUE	Choque Lateral Perdida De Pista Choque Posterior Estrellamientos Choque Frontal	Choque Lateral Choque Posterior Atropellos Perdida De Pista Estrellamientos	Choque Lateral Atropellos Perdida De Pista Choque Posterior Estrellamientos	Choque Lateral Perdida De Pista Choque Posterior Estrellamientos Atropellos	Choque Lateral Atropellos Choque Posterior Perdida De Pista Estrellamientos	Perdida De Pista Choque Lateral Choque Posterior Estrellamientos Choque Frontal
CAUSA PROBABLE	C14 C23 C11 C18 C12	C23 C09 C11 C19 C06	C09 C23 C11 C19 C06	C09 C06 C23 C11 C19	C23 C09 C19 C11 C12	C14 C11 C23 C18 C12
MES	Enero Junio Marzo Julio Abril	Enero Abril Agosto Junio Mayo	Agosto Junio Diciembre Febrero Noviembre	Octubre Agosto Diciembre Noviembre Junio	Marzo Enero Junio Febrero Mayo	Enero Octubre Agosto Julio Junio
DIA	Lunes Martes Jueves Miercoles Viernes	Jueves Miercoles Viernes Martes Sabado	Jueves Viernes Martes Miercoles Sabado	Domingo Sabado Viernes Jueves	Lunes Martes Miercoles Domingo	Domingo Sabado Viernes Jueves
PERIODO	De 06H00 A 06H59 De 22H00 A 22H59 De 20H00 A 20H59 De 18H00 A 18H59 De 19H00 A 19H59	De 07H00 A 07H59 De 15H00 A 15H59 De 17H00 A 17H59 De 18H00 A 18H59 De 16H00 A 16H59	De 16H00 A 16H59 De 20H00 A 20H59 De 18H00 A 18H59 De 17H00 A 17H59 De 14H00 A 14H59	De 07H00 A 07H59 De 08H00 A 08H59 De 06H00 A 06H59 De 00H00 A 00H59 De 16H00 A 16H59	De 14H00 A 14H59 De 15H00 A 15H59 De 16H00 A 16H59 De 20H00 A 20H59 De 18H00 A 18H59	De 06H00 A 06H59 De 13H00 A 13H59 De 02H00 A 02H59 De 00H00 A 00H59 De 16H00 A 16H59
CANTON	Daule Guayaquil Naranjal San Jacinto De Yaguachi Duran	Guayaquil Milagro Duran Daule Samborondon	Guayaquil Milagro Samborondon Duran Daule	Guayaquil Milagro Duran Daule Samborondon	Guayaquil Milagro Duran Samborondon Daule	Daule San Jacinto De Yaguachi Guayaquil Naranjal Duran

4.1.2.2. Pichincha

Utilizando la estrategia aglomerativa del modelo de Agrupación Jerárquico, sobre el conjunto de datos se obtuvo que el número óptimo de conglomerados para Guayas es $k = 5$, en la Figura 4-9 se muestra el dendrograma obtenido.

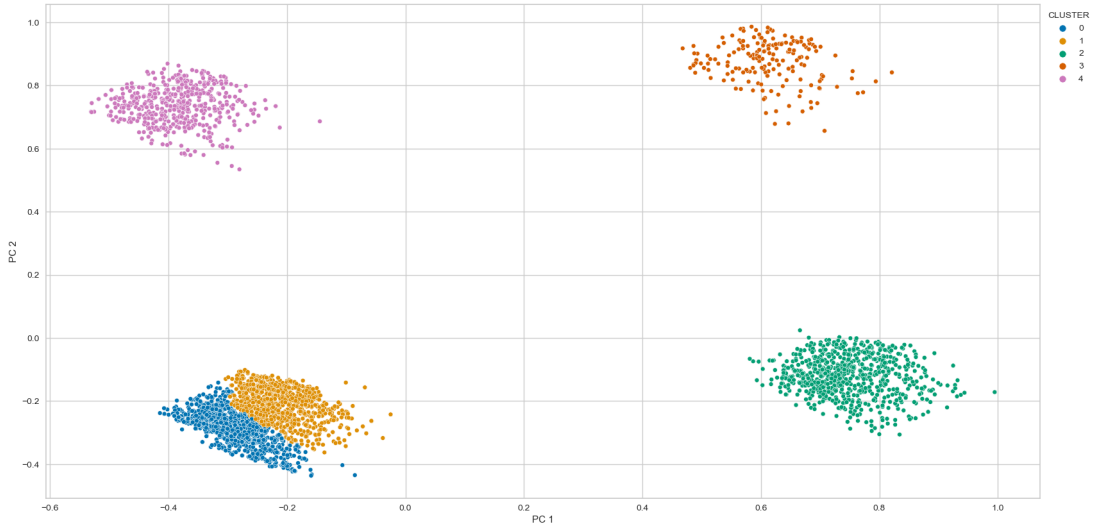
Figura 4-9 Dendrograma para el valor óptimo de k en Agrupación Jerárquica, Pichincha



Fuente (Maza. E, 2022)

Por lo tanto, $k = 5$ óptimo de clústeres ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente Figura 4-8 muestra la visualización de clústeres para la provincia de Pichincha.

Figura 4-10 Agrupación Jerárquica clústeres, Pichincha



Fuente (Maza. E, 2022)

La Tabla 4-5, cataloga las principales características para los clústeres obtenidos con el modelo Agrupación Jerárquica en la provincia de Pichincha.

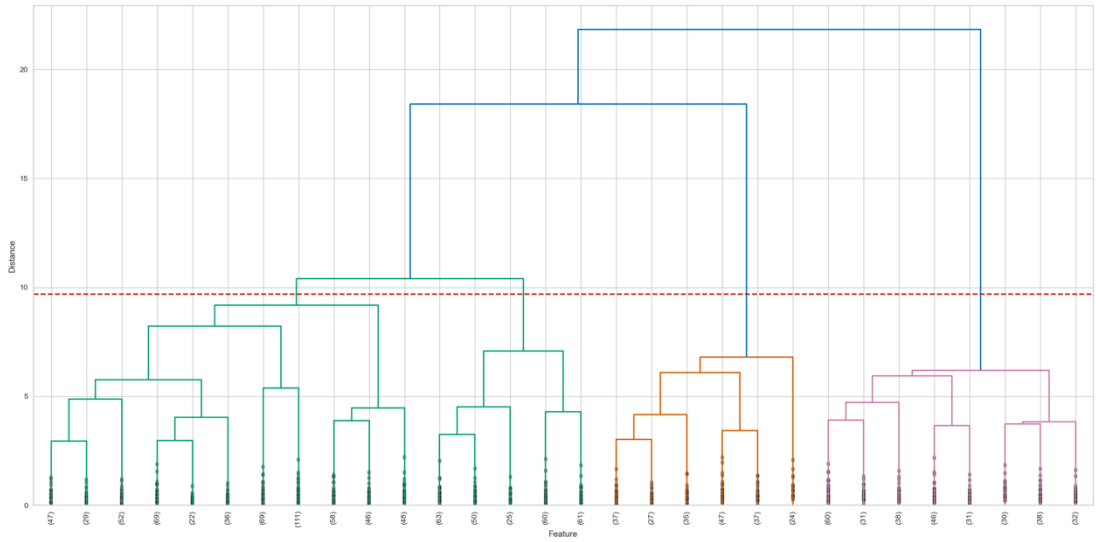
Tabla 4-5 Principales características por clúster en la provincia de Pichincha, Agrupación Jerárquica

	CLUSTER				
CARACTERÍSTICAS	**0**	**1**	**2**	**3**	**4**
ZONA	Urbana	Urbana	Rural	Rural	Urbana
SEXO	Hombre	Mujer	Hombre	Mujer	Hombre
FERIADO	No Si	No Si	No Si	No Si	No Si
TIPO VEHÍCULO	Automovil Motocicleta Camioneta Camion Bus	Automovil Motocicleta Camioneta Bus Camion	Automovil Motocicleta Camion Camioneta Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Vehiculo Deportivo Utilitario Bus	Automovil Bus Camioneta Camion Motocicleta
TIPO CHOQUE	Choque Lateral Estrellamientos Choque Posterior Atropellos Otros	Choque Lateral Estrellamientos Atropellos Choque Posterior Colision	Choque Lateral Estrellamientos Choque Posterior Perdida De Carril Choque Frontal	Choque Lateral Perdida De Carril Estrellamientos Atropellos Choque Frontal	Choque Lateral Estrellamientos Choque Posterior Atropellos Colision
CAUSA PROBABLE	C06 C09 C23 C10 C25	C23 C09 C06 C10 C25	C14 C06 C09 C23 C10	C14 C06 C09 C23 C10	C06 C23 C09 C11 C25
MES	Diciembre Octubre Noviembre Septiembre Agosto	Septiembre Julio Agosto Junio Diciembre	Agosto Septiembre Marzo Enero Julio	Enero Marzo Agosto Febrero Junio	Enero Marzo Febrero Abril Mayo
DIA	Sabado Domingo Viernes Jueves Miercoles	Viernes Sabado Domingo Jueves Miercoles	Domingo Sabado Viernes Jueves Miercoles	Sabado Domingo Viernes Jueves Lunes	Viernes Miercoles Lunes Martes Sabado
PERIODO	De 18H00 A 18H59 De 14H00 A 14H59 De 12H00 A 12H59 De 19H00 A 19H59 De 21H00 A 21H59	De 08H00 A 08H59 De 15H00 A 15H59 De 21H00 A 21H59 De 18H00 A 18H59 De 13H00 A 13H59	De 19H00 A 19H59 De 16H00 A 16H59 De 20H00 A 20H59 De 17H00 A 17H59 De 06H00 A 06H59	De 12H00 A 12H59 De 15H00 A 15H59 De 18H00 A 18H59 De 19H00 A 19H59 De 21H00 A 21H59	De 16H00 A 16H59 De 17H00 A 17H59 De 15H00 A 15H59 De 19H00 A 19H59 De 14H00 A 14H59
CANTON	Quito Ruminiahui Mejia Cayambe Pedro Moncayo	Quito Ruminiahui Cayambe	Quito Mejia Cayambe Pedro Moncayo Ruminiahui	Quito Mejia Pedro Moncayo Cayambe Ruminiahui	Quito Ruminiahui Cayambe Mejia Pedro Moncayo

4.1.2.3. Manabí

Utilizando la estrategia aglomerativa del modelo de Agrupación Jerárquico, sobre el conjunto de datos se obtuvo que el número óptimo de conglomerados para Manabí es $k = 4$, en la Figura 4-11 se muestra el dendrograma obtenido.

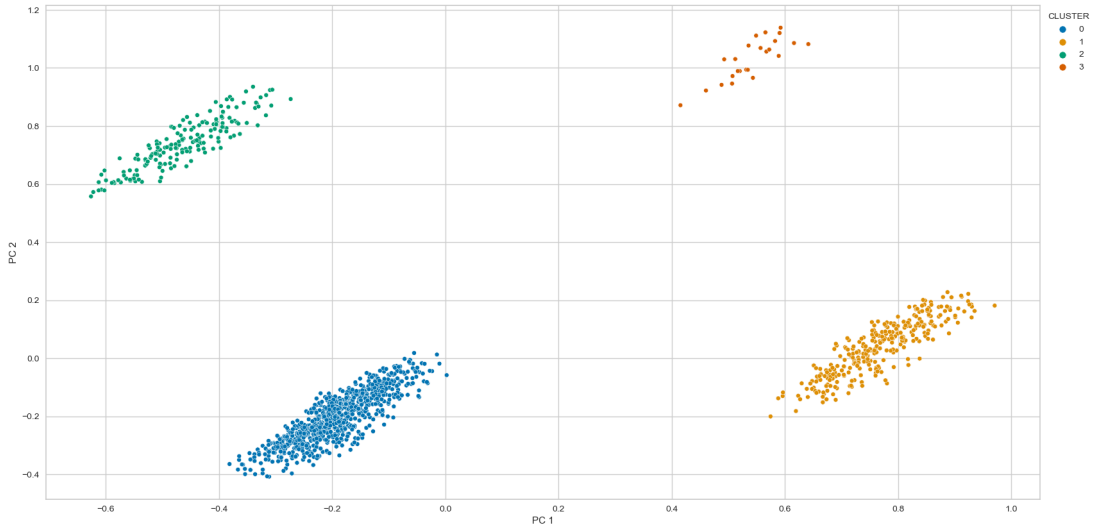
Figura 4-11 Dendrograma para el valor óptimo de k en Agrupación Jerárquica, Manabí



Fuente (Maza. E, 2022)

Por lo tanto, $k = 4$ óptimo de clústeres ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente muestra la visualización de clústeres para la provincia de Manabí.

Figura 4-12 Agrupación Jerárquica clústeres, Manabí



Fuente (Maza. E, 2022)

La Tabla 4-6, cataloga las principales características para los clústeres obtenidos con el modelo Agrupación Jerárquica en la provincia de Manabí.

Tabla 4-6 Principales características por clúster en la provincia de Manabí, Agrupación Jerárquica

	CLUSTER			
CARACTERÍSTICAS	**0**	**1**	**2**	**3**
ZONA	Urbana	Rural	Urbana	Rural
SEXO	Hombre	Hombre	Mujer	Mujer
FERIADO	No Si	No Si	No Si	No Si
TIPO VEHÍCULO	Automovil Motocicleta Camioneta Camion Bicicleta	Motocicleta Automovil Camioneta Camion Vehiculo Deportivo Utilitario	Automovil Motocicleta Camioneta Vehiculo Deportivo Utilitario Bicicleta	Automovil Motocicleta Camioneta Vehiculo Deportivo Utilitario
TIPO CHOQUE	Choque Lateral Choque Posterior Choque Frontal Rozamientos Estrellamientos	Choque Lateral Perdida De Pista Choque Posterior Estrellamientos Choque Frontal	Choque Lateral Choque Posterior Estrellamientos Rozamientos Atropellos	Choque Lateral Perdida De Pista Choque Frontal Estrellamientos Perdida De Carril
CAUSA PROBABLE	C14 C25 C19 C11 C12	C14 C23 C11 C12 C18	C14 C25 C11 C12 C19	C14 C23 C04 C09 C01
MES	Diciembre Agosto Noviembre Septiembre Abril	Enero Marzo Junio Octubre Agosto	Junio Septiembre Julio Agosto Diciembre	Enero Abril Febrero Mayo Agosto
DIA	Sabado Viernes Domingo Jueves Miercoles	Domingo Sabado Jueves Viernes Martes	Domingo Viernes Jueves Martes Lunes	Domingo Sabado Viernes Lunes Miercoles
PERIODO	De 17H00 A 17H59 De 13H00 A 13H59 De 19H00 A 19H59 De 15H00 A 15H59 De 20H00 A 20H59	De 17H00 A 17H59 De 20H00 A 20H59 De 19H00 A 19H59 De 12H00 A 12H59 De 18H00 A 18H59	De 14H00 A 14H59 De 11H00 A 11H59 De 16H00 A 16H59 De 19H00 A 19H59 De 17H00 A 17H59	De 17H00 A 17H59 De 18H00 A 18H59 De 09H00 A 09H59 De 15H00 A 15H59 De 19H00 A 19H59
CANTON	Manta Portoviejo El Carmen Jipijapa Pedernales	Portoviejo Montecristi Rocafuerte Chone Manta	Manta Portoviejo El Carmen Jaramijo Jipijapa	Jipijapa Montecristi Pedernales Portoviejo Santa Ana

4.1.3. DBSCAN

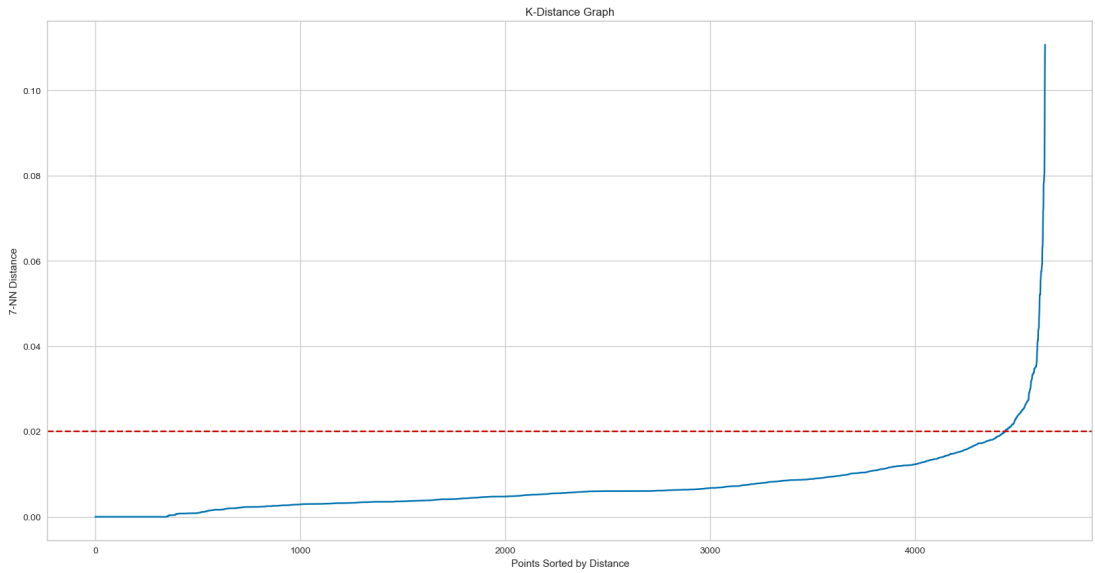
Utilizando el algoritmo DBSCAN se obtuvo los siguientes resultados, cabe resaltar que el modelo DBSCAN está basada en la densidad.

4.1.3.1. Guayas

Utilizando el modelo DBSCAN se determinó los parámetros $eps = 0.02$ y basados en las reglas se ha determinado $min\_samples = 4$.

-

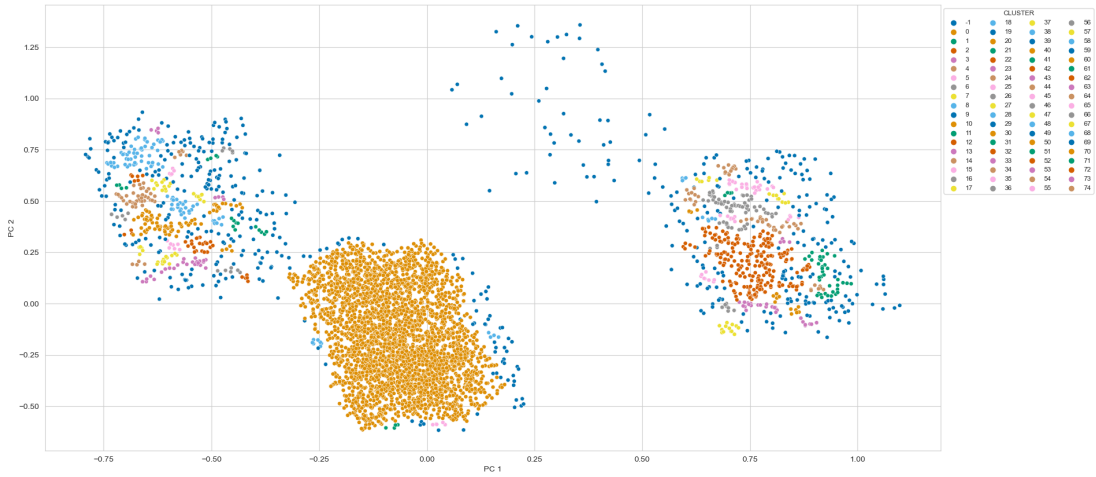
Figura 4-13 Valor óptimo de épsilon en DBSCAN, Guayas



Fuente (Maza. E, 2022)

Por lo tanto, $eps = 0.02$ y $min\_sample = 4$ óptimos, ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente Figura 4-14 muestra la visualización de clústeres para la provincia de Guayas.

Figura 4-14 DBSCAN clústeres, Guayas

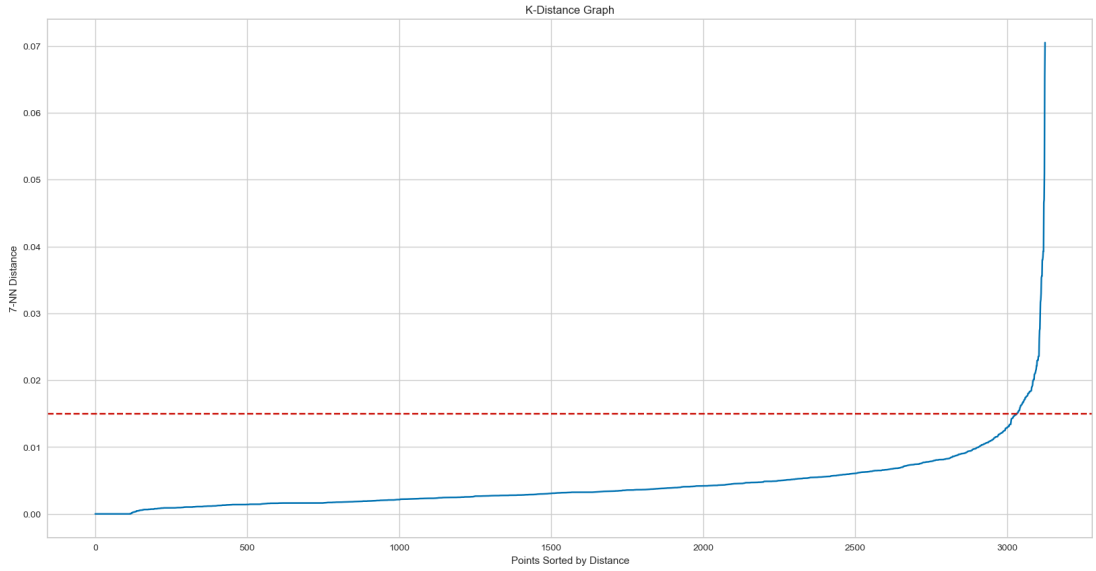


Fuente (Maza. E, 2022)

4.1.3.2. Pichincha

Utilizando el modelo DBSCAN se determinó los parámetros $eps = 0.015$ y basados en las reglas se ha determinado $min\_samples = 4$.

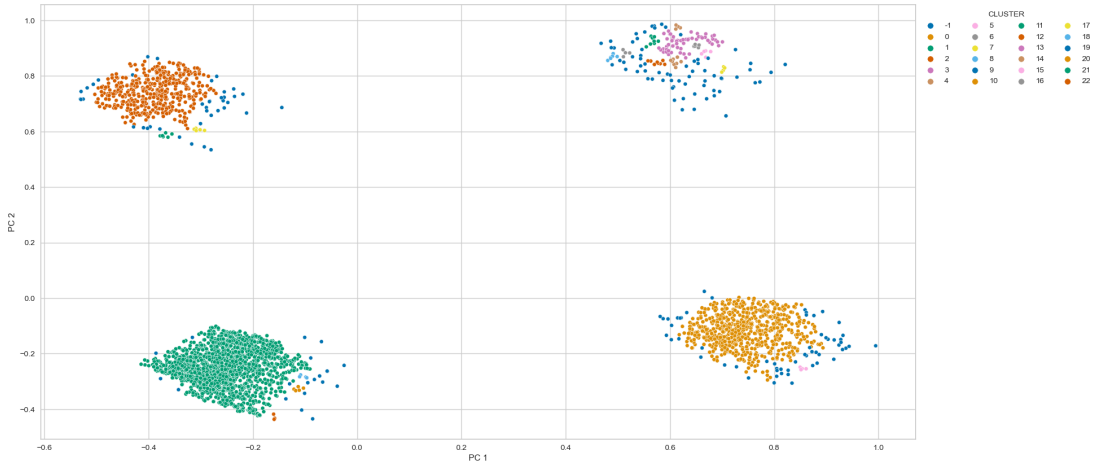
Figura 4-15 Valor óptimo de épsilon en DBSCAN, Pichincha



Fuente (Maza. E, 2022)

Por lo tanto, $eps = 0.015$ y $min\_sample = 4$ óptimos, ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente muestra la visualización de clústeres para la provincia de Guayas.

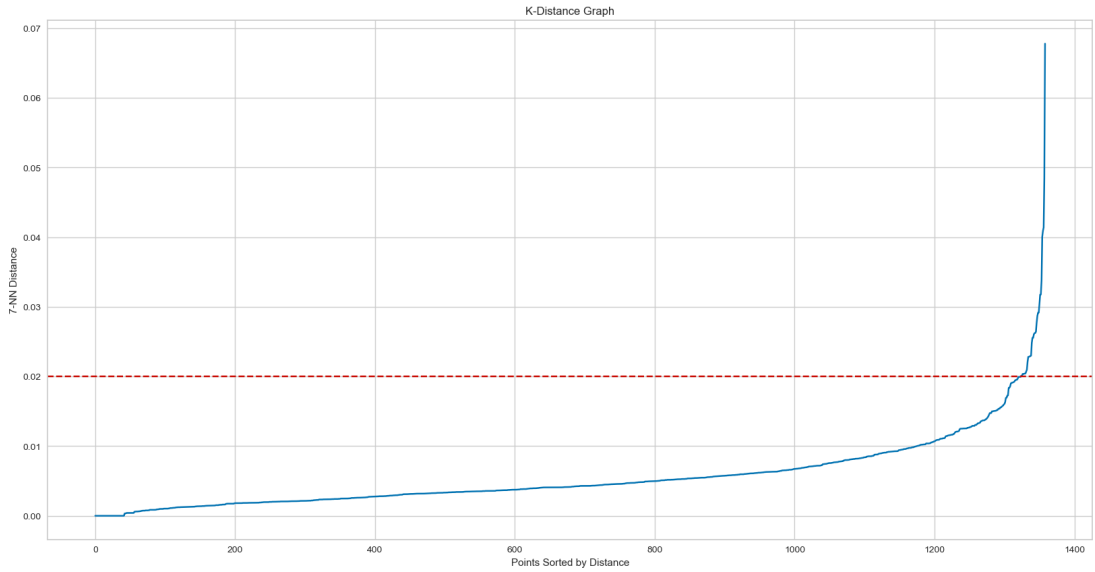
Figura 4-16 DBSCAN clústeres, Pichincha



Fuente (Maza. E, 2022)

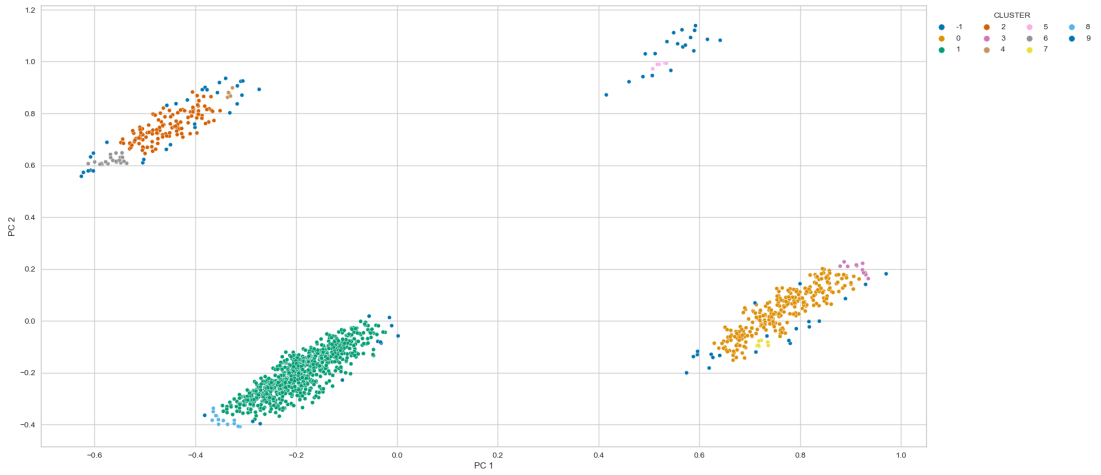
4.1.3.3. Manabí

Utilizando el modelo DBSCAN se determinó los parámetros $eps = 0.02$ y basados en las reglas se ha determinado $min\_samples = 4$.



Fuente (Maza. E, 2022)

Por lo tanto, $eps = 0.02$ y $min\_sample = 4$ óptimos, ha permitido obtener los conglomerados y sus caracterizaciones, la siguiente muestra la visualización de clústeres para la provincia de Manabí.



Fuente (Maza. E, 2022)

4.2. Análisis y Validación de Métricas

Dada la relevancia y caracterizaciones encontradas en los clústeres de las principales provincias que inciden el 63.33 % de los siniestros viales del Ecuador en el año 2021, es de importancia evaluar los resultados obtenidos mediante los índices de validación de agrupamiento.

4.2.1. Guayas

Calinski–Harabasz, el índice mide la relación entre la varianza de cada clúster y la varianza de todos los clústeres. El algoritmo K-means tiene un índice de 880.558 que indica mejor definición y separación de los clústeres. Al igual que ocurre con el coeficiente de Silueta, el índice no funciona adecuadamente con algoritmos basados en la densidad como DBSCAN.

Davies Bouldin, el índice mide la similitud media de cada clúster con su conglomerado más similar. El modelo de K-means tiene un índice de 2.064 el más bajo en relación con los demás modelos, lo que indica una mejor separación entre los clústeres.

Silhouette, la puntuación obtenida del coeficiente de silueta indica que el algoritmo de K-mean tiene un mejor rendimiento con una puntuación de 0.186. También es importante aclarar que esta puntuación no funciona adecuadamente con los algoritmos basados en la densidad, como DBSCAN.

Por lo tanto, para la provincia de Guayas el modelo de K-means tiene mejor rendimiento en base a los índices analizados, en la Tabla 4-7, se muestra la comparación de las métricas.

Tabla 4-7 Métricas de evaluación de modelos, Guayas

Modelo	Índice Calinski	Índice Davies	Silhouette
K-means	880.558	2.064	0.186
Hierarchical Clustering	767.174	2.437	0.123
DBSCAN	31.740	3.315	-0.190

4.2.2. Pichincha

Calinski–Harabasz, El modelo K-means tiene un índice de 789.294 que indica mejor definición y separación de los clústeres. Al igual que ocurre con el coeficiente de Silueta, el índice no funciona adecuadamente con algoritmos basados en la densidad como DBSCAN.

Davies Bouldin, el modelo de K-means tiene un índice de 1.599 el más bajo en relación con los demás modelos, lo que indica una mejor separación entre los clústeres.

Silhouette, el modelo de K-mean tiene un mejor rendimiento con una puntuación de 0.236. También es importante aclarar que esta puntuación no funciona adecuadamente con los algoritmos basados en la densidad, como DBSCAN.

Por lo tanto, para la provincia de Pichincha el modelo de K-means tiene mejor rendimiento en base a los índices analizados, en la Tabla 4-8, se muestra la comparación de las métricas.

Tabla 4-8 Métricas de evaluación de modelos, Pichincha

Modelo	Índice Calinski	Índice Davies	Silhouette
K-means	789.924	1.599	0.236
Hierarchical Clustering	779.631	1.636	0.231
DBSCAN	106.173	2.223	0.035

4.2.3. Manabí

Calinski–Harabasz, El modelo K-means y Agrupamiento Jerárquico tiene un índice de 303.576 que indica mejor definición y separación de los clústeres. Al igual

que ocurre con el coeficiente de Silueta, el índice no funciona adecuadamente con algoritmos basados en la densidad como DBSCAN.

Davies Bouldin, el modelo de K-means y Agrupamiento Jerárquico tiene un índice de 1.334 el más bajo en relación con los demás modelos, lo que indica una mejor separación entre los clústeres.

Silhouette, el modelo de K-mean y Agrupamiento Jerárquico tiene un mejor rendimiento con una puntuación de 0.319. También es importante aclarar que esta puntuación no funciona adecuadamente con los algoritmos basados en la densidad, como DBSCAN.

Por lo tanto, para la provincia de Manabí el modelo de K-means y Agrupamiento Jerárquico tiene mejor rendimiento en base a los índices analizados, en la Tabla 4-9, se muestra la comparación de las métricas.

Tabla 4-9 Métricas de evaluación de modelos, Manabí

Modelo	Índice Calinski	Índice Davies	Silhouette
K-means	303.576	1.334	0.319
Hierarchical Clustering	303.576	1.334	0.319
DBSCAN	87.169	2.256	0.030

Capitulo 5

5. Conclusiones y Recomendaciones

Finalmente, en este capítulo se realiza una recapitulación de los objetivos de la presente investigación aplicada y se exponen las conclusiones y recomendaciones derivadas de los modelos aplicados el conjunto de datos de siniestros viales ocurridos en Ecuador para el año 2021.

5.1. Conclusiones

De acuerdo con los modelos empleados en la presente investigación aplicada utilizando técnicas de aprendizaje no supervisado, para identificar clústeres y patrones que mayormente contribuyen a los casos de siniestralidad vial en Ecuador para el año 2021, de forma que se logre caracterizar los eventos más recurrentes, proporcionando información relevante a los organismos de control vial, se concluye lo siguiente:

- Dado que los siniestros ocurren en todo el territorio ecuatoriano y acorde a las inferencias realizadas en las etapas de *Comprensión de los Datos* y *Preparación de los Datos* la presente investigación aplicada se han enfocado en las principales provincias que inciden en un 63.33 % de accidentes de tránsito. La provincia de Guayas reportó 7.758 siniestros y representa el 36.33 % de los siniestros registrados, Pichincha reporto 3.807 siniestros y representa el 17.83 % de siniestros registrados y Manabí que reporto 1.957 siniestros que representa el 9.17% de siniestros en el año 2021.
- El modelo que presenta las mejores métricas e índices de evaluación para identificar los clústeres y posteriormente caracterizar los conglomerados encontrados es el modelo de K-mean. La provincia de Guayas obtuvo los siguientes resultados: *Coeficiente de Silhouette* de 0.186, índice de *Calinski–Harabasz* de 880.558 y el índice de *Davies Bouldin* de 2.064, Pichincha obtuvo los siguientes resultados: *Coeficiente de Silhouette* de 0.236, índice de *Calinski–Harabasz* de 789.924 y el índice de *Davies*

Bouldin de 1.599 y Manabí obtuvo los siguientes resultados: *Coeficiente de Silhouette* de 0.319, índice de *Calinski–Harabasz* de 303.576 y el índice de *Davies Bouldin* de 1.334.

- Los modelos de aprendizaje no supervisados *K-means*, *Agrupación Jerárquica* y *DBSCAN* aplicados sobre el conjunto de datos ayudaron a identificar los conglomerados con mayor incidencia en los siniestros viales. Asimismo, se caracterizó los eventos más recurrentes basándose en las propiedades de cada uno de los clústeres. Análisis que permite focalizar soluciones por las entidades de control, haciendo posible la disminución de siniestros viales.
- Aunque *K-means* ha definido 5 clústeres para la provincia de Guayas, se han caracterizado las siguientes propiedades que mayormente incurren en siniestros como: suceden mayormente en zonas urbanas, con mayor frecuencia en los cantones de Daule, Guayaquil, Naranjal, San Jacinto de Yaguachi, Duran y Samborondón, en su mayoría son ocasionadas por el sexo masculino. Adicionalmente las principales causas probables son: C23, C09 y C14, por lo regular los tipos de vehículos involucrados son automóviles, motocicletas y camiones, se registran esencialmente como choques laterales, perdida de pista u atropellos. Por último, los siniestros registrados ocurren principalmente en los meses de agosto, marzo, enero y diciembre, comúnmente en los sábados, domingo y lunes, en los periodos de 07H-07H59, 15H00-15H59 y 08H00-08H59.
- Similarmente *K-means* ha definido 5 clústeres para la provincia de Pichincha. Sin embargo, tienen diferentes caracterizaciones y las propiedades que mayormente incurren en los siniestros son: suceden mayormente en zonas urbanas, con mayor frecuencia en los cantones de Quito, Mejía, Rumiñahui y Cayambe, en su mayoría son ocasionadas por el sexo masculino. Adicionalmente las principales causas probables son: C14, C09, C06 y C23, por lo regular los tipos de vehículos involucrados son automóviles, motocicletas, camiones y buses, se registran esencialmente

como choques laterales, estrellamientos y choques posteriores. Por último, los siniestros registrados ocurren principalmente en los meses de octubre, diciembre, enero y marzo, comúnmente en los sábados, domingo, y viernes, en los periodos de 19H00-19H59, 15H00-15H59 y 21H00-21H59.

- En cuanto a la provincia de Manabí, K-*means* ha definido 4 clústeres, se han caracterizado las subsiguientes propiedades que mayormente incurren en siniestros como: suceden mayormente en zonas urbanas, con mayor frecuencia en los cantones de Manta, Portoviejo, Montecristi y Jipijapa, en su mayoría son ocasionadas por el sexo masculino. Adicionalmente las principales causas probables son: C14, C25, C19 y C23, por lo regular los tipos de vehículos involucrados son automóviles, motocicletas y camioneta, se registran esencialmente como choque lateral o choque posterior. Por último, los siniestros registrados ocurren principalmente en los meses de diciembre, agosto, noviembre y abril, comúnmente en los viernes, sábado, y domingo, en los periodos de 14H00-14H59, 11H00-11H59 y 16H00-16H59.

5.2. Recomendaciones

- Con la finalidad de mejorar la caracterización de los conglomerados, se debe buscar la inclusión de nuevas variables dentro del modelo como: factores climáticos, fallas mecánicas, estado vial, factor de señalización.
- Para trabajos futuros se puede profundizar el análisis con otros modelos de clusterización que permitan mejorar los resultados obtenidos. Adicionalmente, considérese incluir expertos del dominio de acuerdo con el tema de interés u objetivos del negocio.
- Dado que se aplicó los modelos de aprendizaje no supervisado más utilizados en la clusterización de siniestros en las principales provincias, es posible aplicar la misma metodología en las provincias no incluidas y caracterizar los resultados.

6. Referencias

Global Burden of Disease. (2019). Obtenido de Global Burden of Disease Study: https://vizhub.healthdata.org/gbd-results/

Agencia Nacional de Tránsito. (2021). Obtenido de https://www.ant.gob.ec/visor-de-siniestralidad-estadisticas

Metodología CRIPS DM. (2021). Obtenido de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

Bradley, P. S., & Fayyad, U. M. (1998, July). Refining initial points for k-means clustering. In ICML (Vol. 98, pp. 91-99).

Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained k-means clustering. Microsoft Research, Redmond, 20(0), 0.

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303.

Wu, J. (2012). Advances in K-means clustering: a data mining thinking. Springer Science & Business Media.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology, 9(1), 15-24.

Murtagh, F., & Contreras, P. (2011). Methods of hierarchical clustering. arXiv preprint arXiv:1105.0121.

Contreras, P., & Murtagh, F. (2015). Hierarchical clustering. Handbook of Cluster Analysis; Henning, C., Meila, M., Murtagh, F., Rocci, R., Eds, 103-123.

Reddy, M., Makara, V., & Satish, R. U. V. N. (2017). Divisive hierarchical clustering with K-means and agglomerative hierarchical clustering. Int J of Comp Science Trands and Tech (IJCST), 5(5), 5-11.

Deng, D. (2020, September). DBSCAN clustering algorithm based on density. In 2020 7th international forum on electrical engineering and automation (IFEEA) (pp. 949-953). IEEE.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

Rahmah, N., & Sitanggang, I. S. (2016). Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In IOP conference series: earth and environmental science (Vol. 31, No. 1, p. 012012). IoP Publishing.

Malzer, C., & Baum, M. (2020, September). A hybrid approach to hierarchical density-based cluster selection. In 2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI) (pp. 223-228). IEEE.

Dinh, D. T., Fujinami, T., & Huynh, V. N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20 (pp. 1-17). Springer Singapore.

Petrovic, S. (2006, October). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In Proceedings of the 11th Nordic workshop of secure IT systems (Vol. 2006, pp. 53-64). Citeseer.

Sitompul, B. J. D., Sitompul, O. S., & Sihombing, P. (2019, June). Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. In Journal of Physics: Conference Series (Vol. 1235, No. 1, p. 012015). IOP Publishing.

7. Anexo

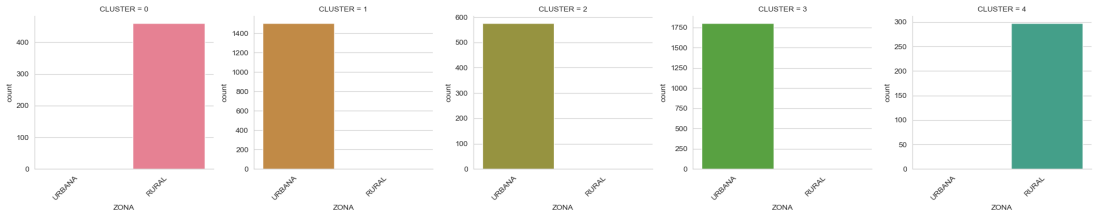
7.1. Diagramas de estudio de clústeres para las principales provincias que inciden el 63.33 % de siniestro viales del Ecuador en el año 2021.

A continuación, se muestra las inferencias más relevantes encontradas en cada uno de los modelos de clústeres aplicados a la provincia de Guayas.

7.1.1. K-means

7.1.1.1. Guayas

Figura 7-1 K-means, Guayas, Clústeres para característica Zona



Fuente (Maza. E, 2022)

Figura 7-2 K-means, Guayas, Clústeres para característica Sexo

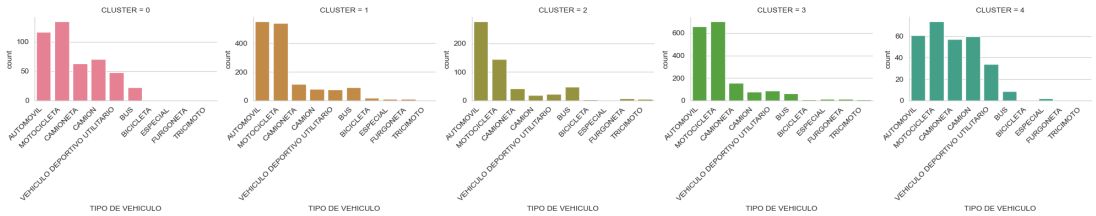
Fuente (Maza. E, 2022)

Figura 7-3 K-means, Guayas, Clústeres para característica Feriado

Fuente (Maza. E, 2022)

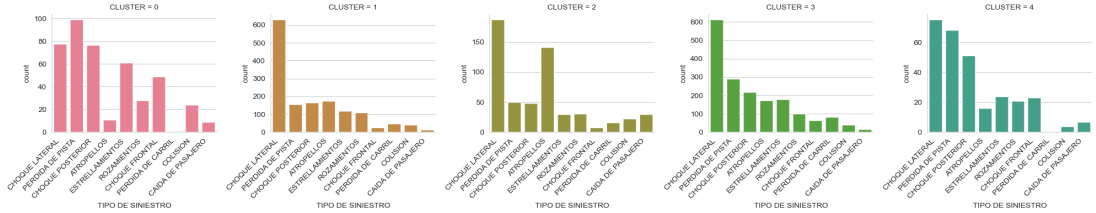
Figura 7-4 K-means, Guayas, Clústeres para característica Tipo de Vehículo

Guayas, Clústeres para característica Feriado



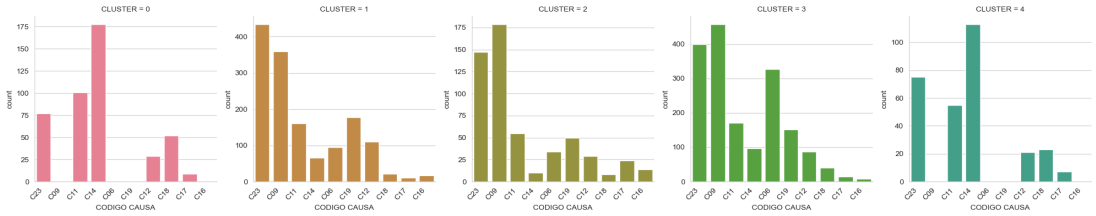
Fuente (Maza. E, 2022)

Figura 7-5 K-means, Guayas, Clústeres para característica Tipo de Siniestro



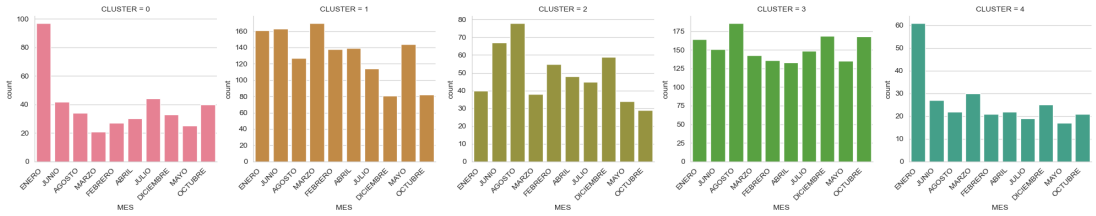
Fuente (Maza. E, 2022)

Figura 7-6 K-means, Guayas, Clústeres para característica Causa Probable



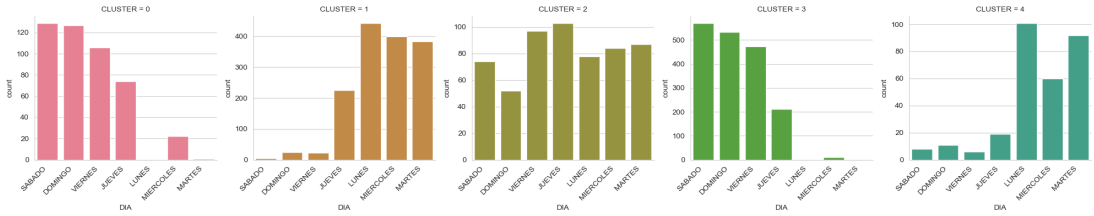
Fuente (Maza. E, 2022)

Figura 7-7 K-means, Guayas, Clústeres para característica Mes



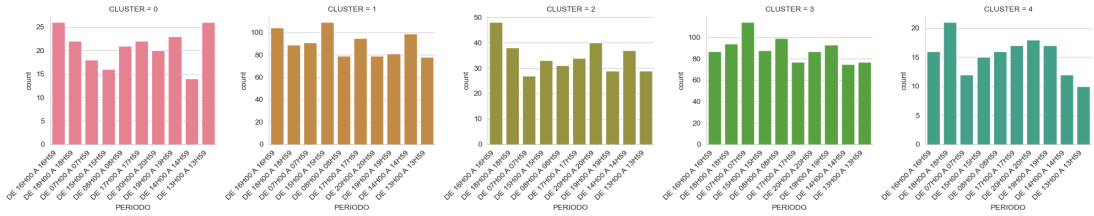
Fuente (Maza. E, 2022)

Figura 7-8 K-means, Guayas, Clústeres para característica Día



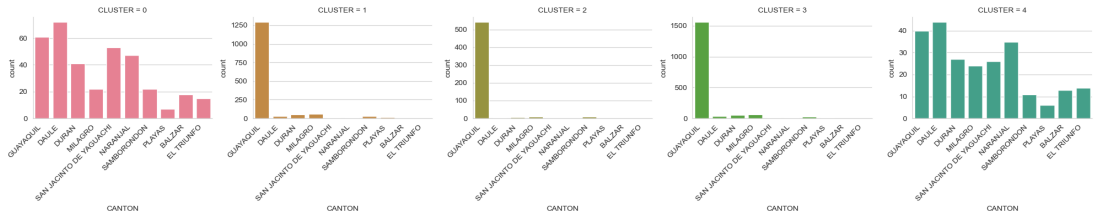
Fuente (Maza. E, 2022)

Figura 7-9 K-means, Guayas, Clústeres para característica Periodo



Fuente (Maza. E, 2022)

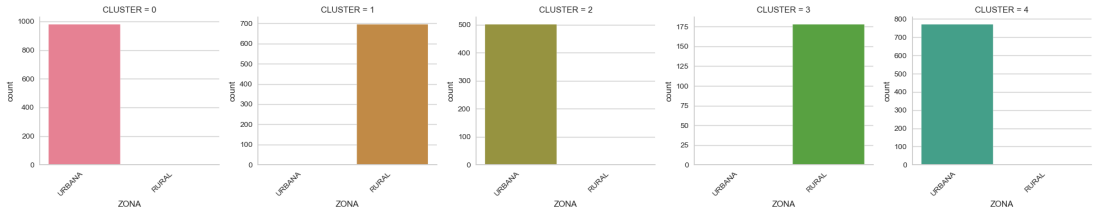
Figura 7-10 K-means, Guayas, Clústeres para característica Cantón



Fuente (Maza. E, 2022)

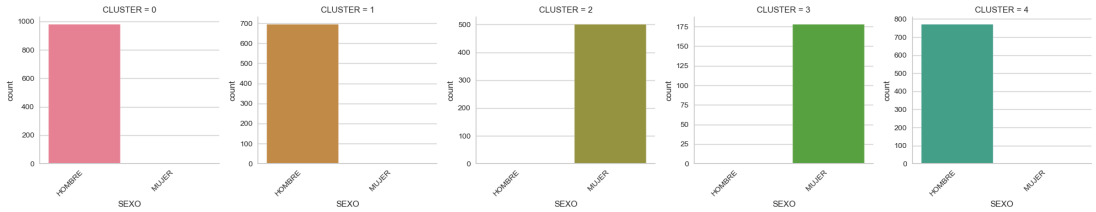
7.1.1.2. Pichincha

Figura 7-11 K-means, Pichincha, Clústeres para característica Zona



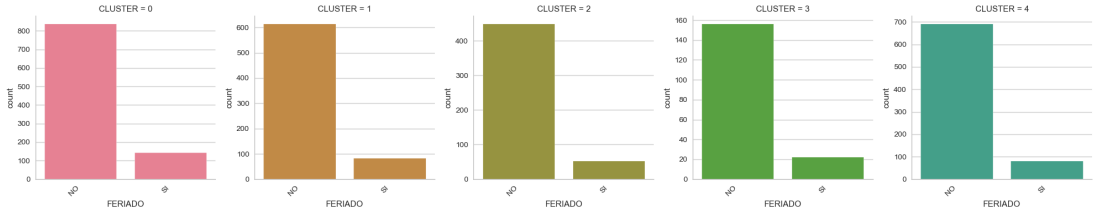
Fuente (Maza. E, 2022)

Figura 7-12 K-means, Pichincha, Clústeres para característica Sexo



Fuente (Maza. E, 2022)

Figura 7-13 K-means, Pichincha, Clústeres para característica Feriado



Fuente (Maza. E, 2022)

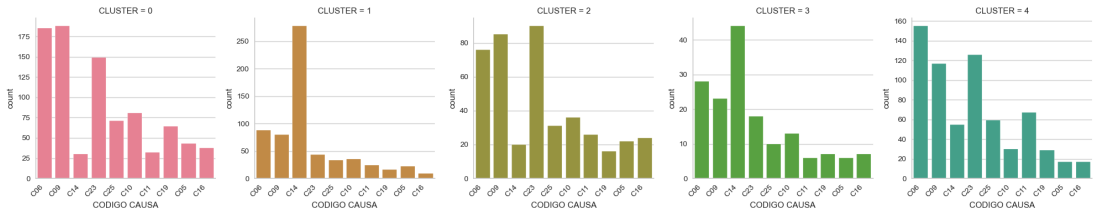
Figura 7-14 K-means, Pichincha, Clústeres para característica Tipo de Vehículo

Fuente (Maza. E, 2022)

Figura 7-15 K-means, Pichincha, Clústeres para característica Tipo de Siniestro

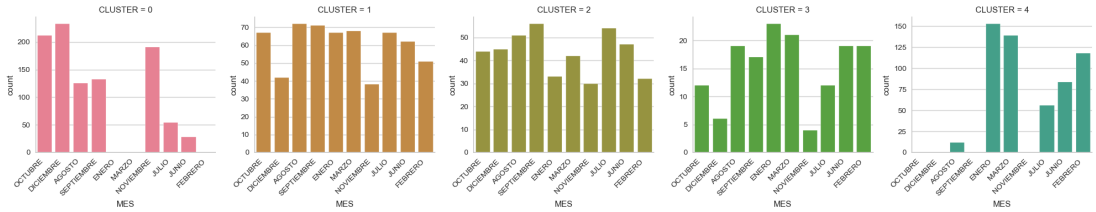
Fuente (Maza. E, 2022)

Figura 7-16 K-means, Pichincha, Clústeres para característica Causa Probable



Fuente (Maza. E, 2022)

Figura 7-17 K-means, Pichincha, Clústeres para característica Mes

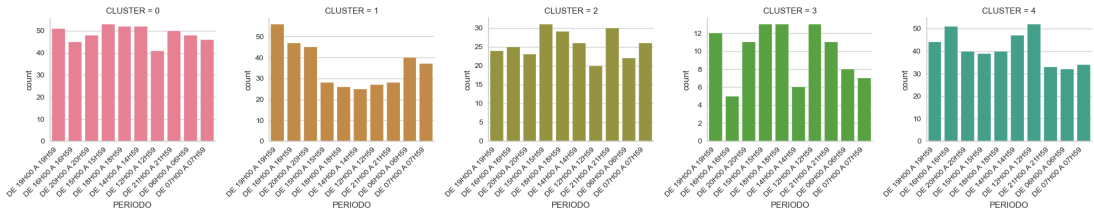


Fuente (Maza. E, 2022)

Figura 7-18 K-means, Pichincha, Clústeres para característica Día

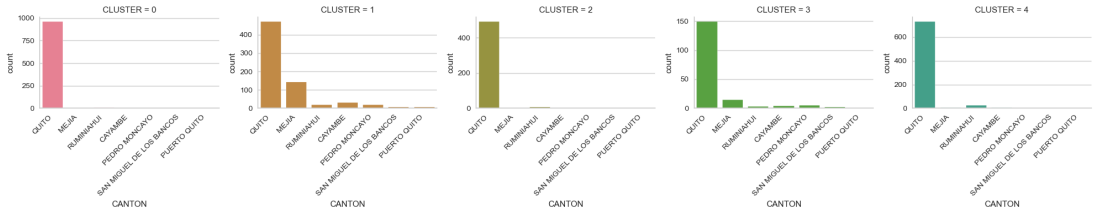
Fuente (Maza. E, 2022)

Figura 7-19 K-means, Pichincha, Clústeres para característica Periodo



Fuente (Maza. E, 2022)

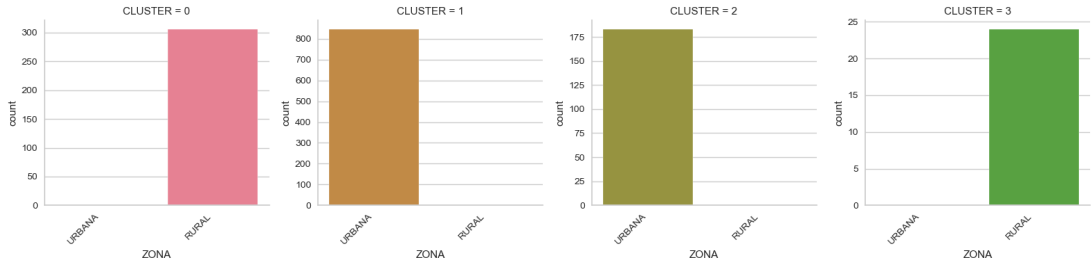
Figura 7-20 K-means, Pichincha, Clústeres para característica Cantón



Fuente (Maza. E, 2022)

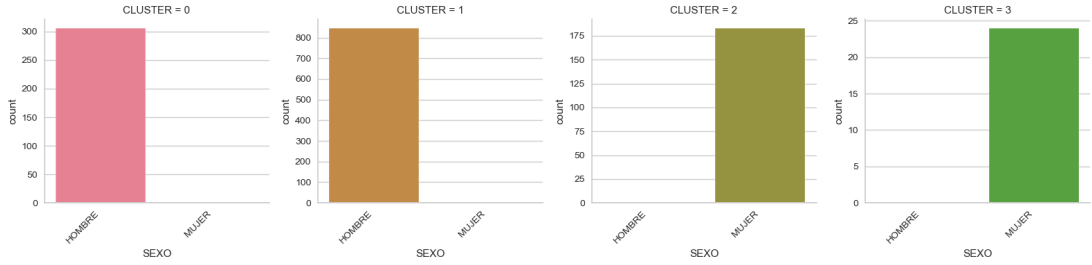
7.1.1.3. Manabí

Figura 7-21 K-means, Manabí, Clústeres para característica Zona



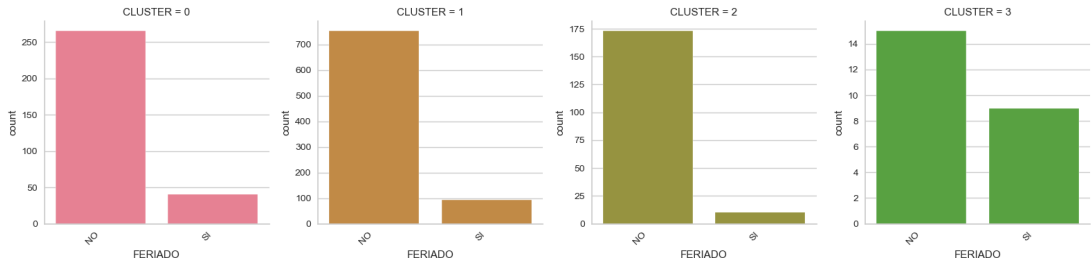
Fuente (Maza. E, 2022)

Figura 7-22 K-means, Manabí, Clústeres para característica Sexo



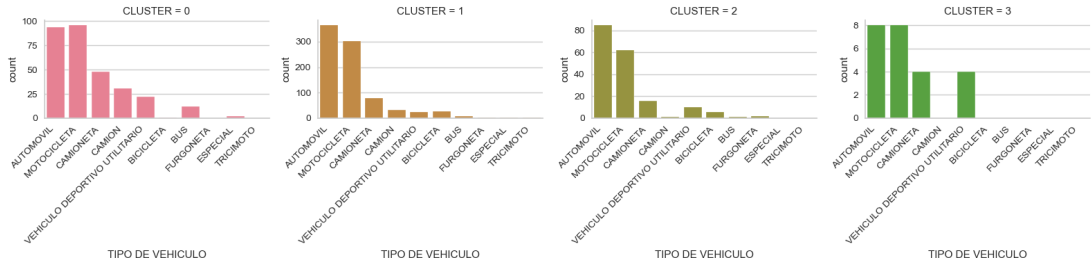
Fuente (Maza. E, 2022)

Figura 7-23 K-means, Manabí, Clústeres para característica Feriado



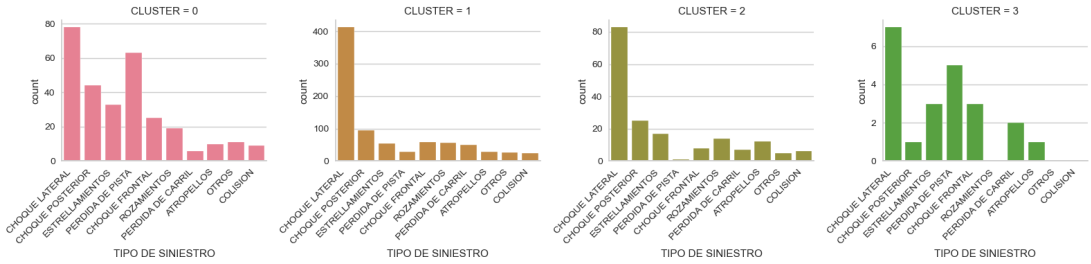
Fuente (Maza. E, 2022)

Figura 7-24 K-means, Manabí, Clústeres para característica Tipo de Vehículo



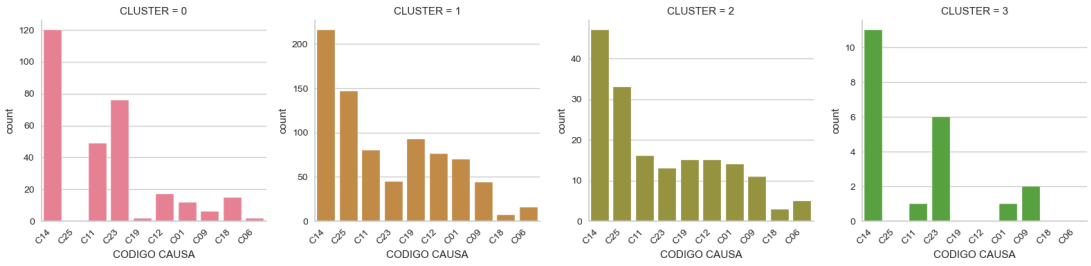
Fuente (Maza. E, 2022)

Figura 7-25 K-means, Manabí, Clústeres para característica Tipo de Siniestro

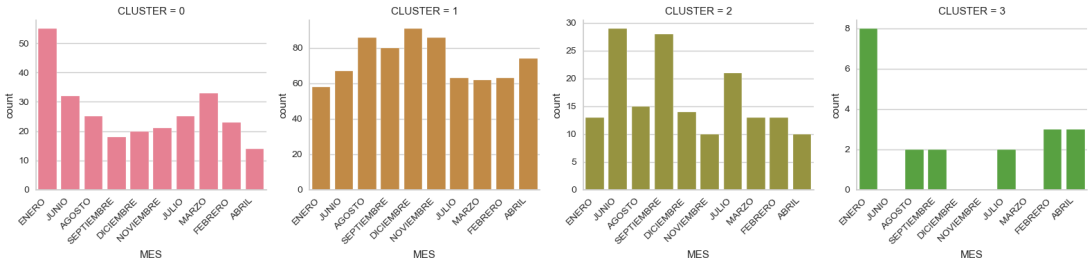


Fuente (Maza. E, 2022)

Figura 7-26 K-means, Manabí, Clústeres para característica Causa Probable

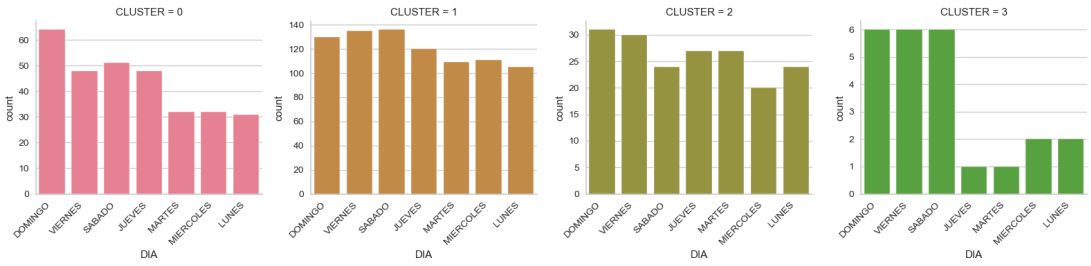


Fuente (Maza. E, 2022)



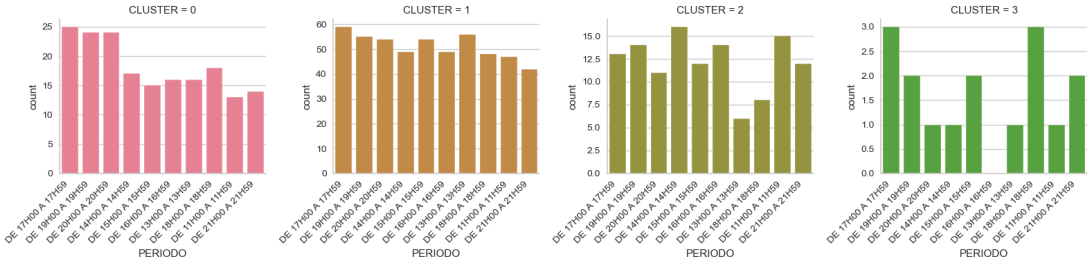
Fuente (Maza. E, 2022)

Figura 7-27 K-means, Manabí, Clústeres para característica Día



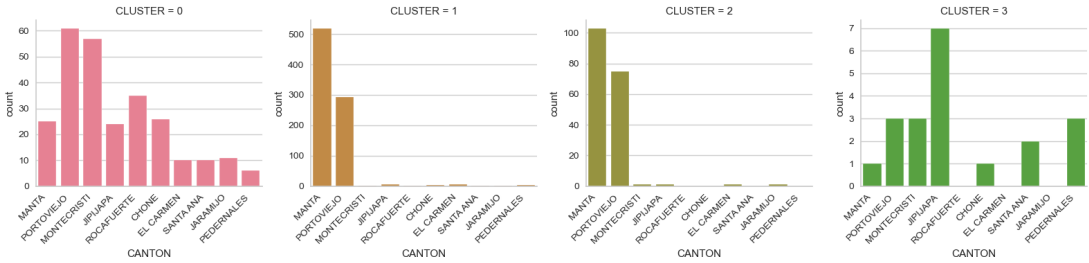
Fuente (Maza. E, 2022)

Figura 7-28 K-means, Manabí, Clústeres para característica Periodo



Fuente (Maza. E, 2022)

Figura 7-29 K-means, Manabí, Clústeres para característica Cantón



Fuente (Maza. E, 2022)

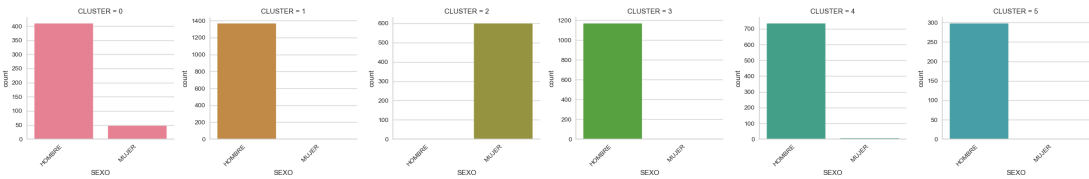
7.1.2. Agrupamiento Jerárquico

7.1.2.1. Guayas

Figura 7-30 Agrupamiento Jerárquico, Guayas, Clústeres para característica Zona

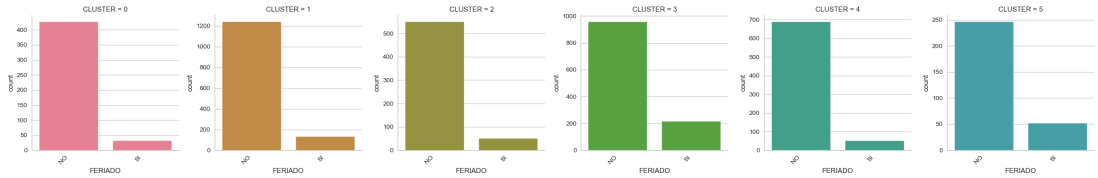
Fuente (Maza. E, 2022)

Figura 7-31 Agrupamiento Jerárquico, Guayas, Clústeres para característica Sexo



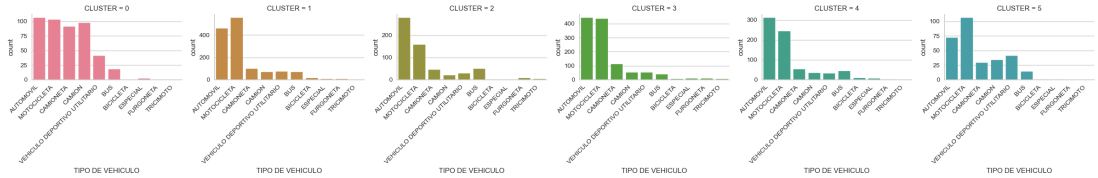
Fuente (Maza. E, 2022)

Figura 7-32 Agrupamiento Jerárquico, Guayas, Clústeres para característica Feriado



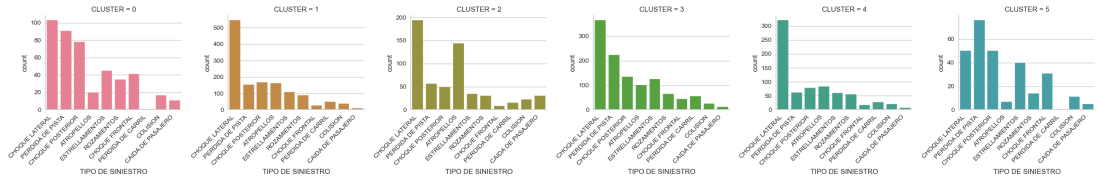
Fuente (Maza. E, 2022)

Figura 7-33 Agrupamiento Jerárquico, Guayas, Clústeres para característica Tipo de Vehículo



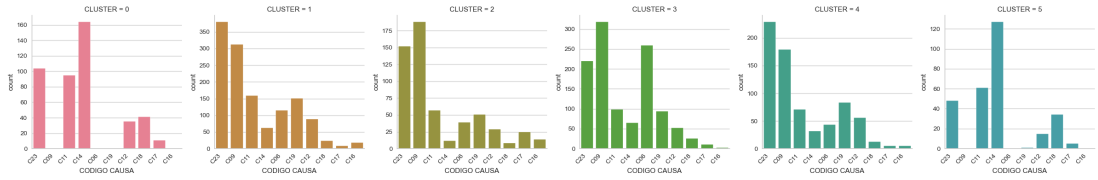
Fuente (Maza. E, 2022)

Figura 7-34 Agrupamiento Jerárquico, Guayas, Clústeres para característica Tipo de Siniestro



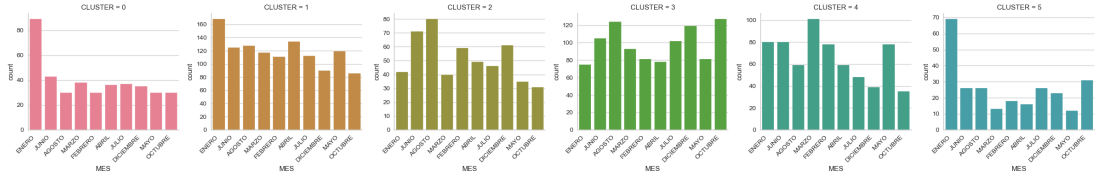
Fuente (Maza. E, 2022)

Figura 7-35 Agrupamiento Jerárquico, Guayas, Clústeres para característica Causa Probable



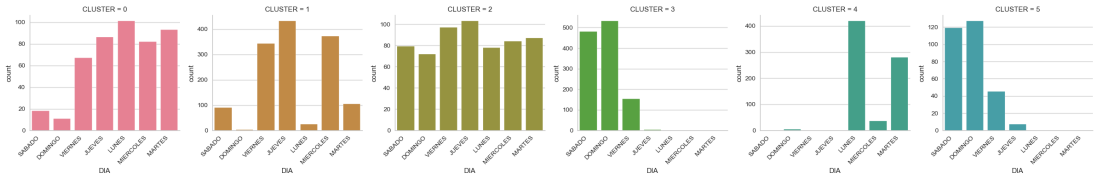
Fuente (Maza. E, 2022)

Figura 7-36 Agrupamiento Jerárquico, Guayas, Clústeres para característica Mes



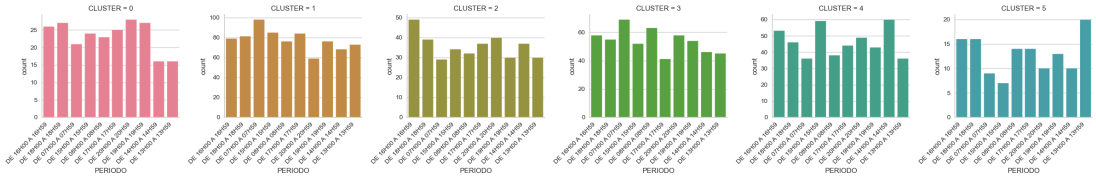
Fuente (Maza. E, 2022)

Figura 7-37 Agrupamiento Jerárquico, Guayas, Clústeres para característica Día



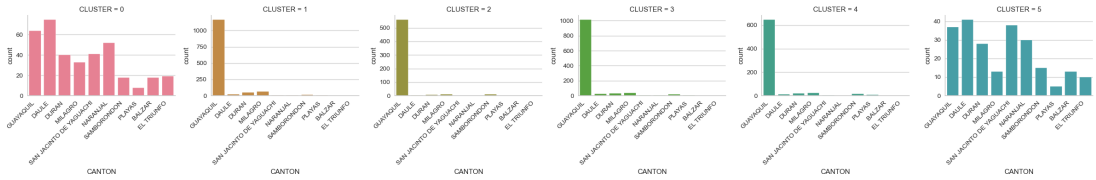
Fuente (Maza. E, 2022)

Figura 7-38 Agrupamiento Jerárquico, Guayas, Clústeres para característica Periodo



Fuente (Maza. E, 2022)

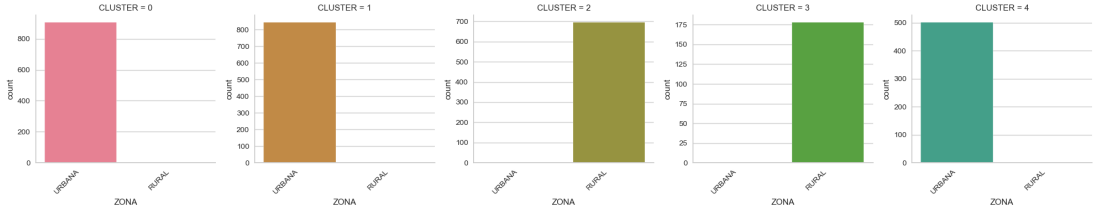
Figura 7-39 Agrupamiento Jerárquico, Guayas, Clústeres para característica Cantón



Fuente (Maza. E, 2022)

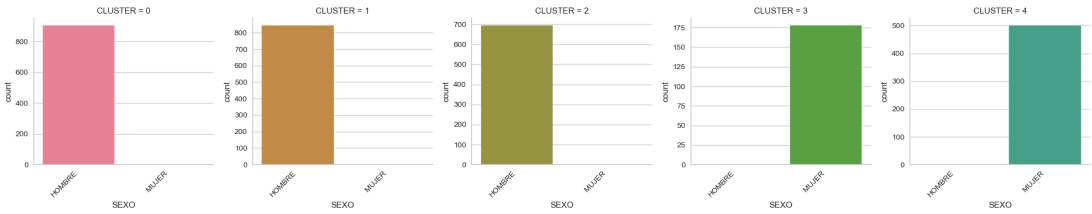
7.1.2.2. Pichincha

Figura 7-40 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Zona



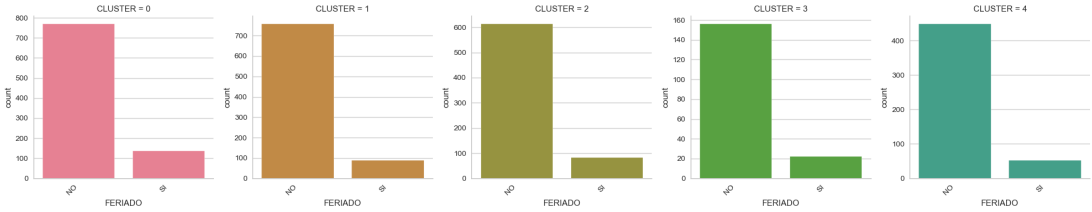
Fuente (Maza. E, 2022)

Figura 7-41 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Sexo



Fuente (Maza. E, 2022)

Figura 7-42 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Feriado



Fuente (Maza. E, 2022)

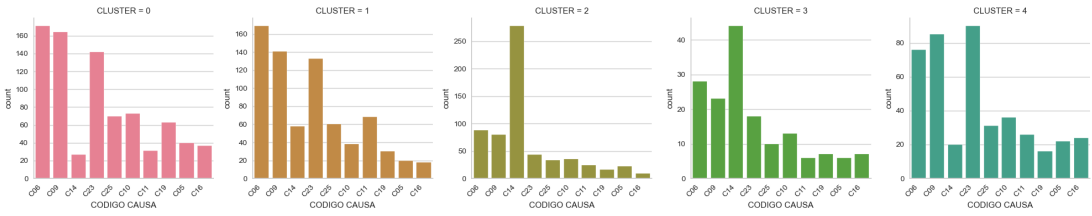
Figura 7-43 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Tipo de Vehículo

Fuente (Maza. E, 2022)

Figura 7-44 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Tipo de Siniestro

Fuente (Maza. E, 2022)

Figura 7-45 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Causa Probable



Fuente (Maza. E, 2022)

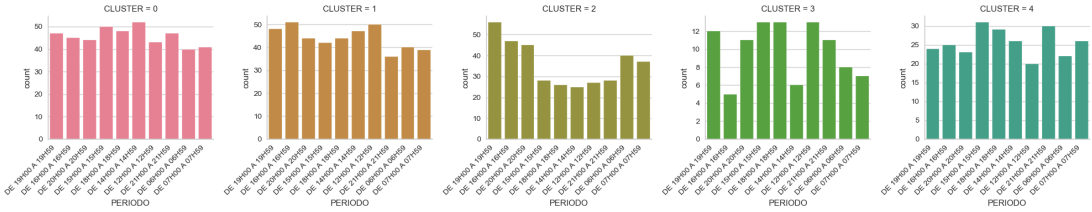
Figura 7-46 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Mes

Fuente (Maza. E, 2022)

Figura 7-47 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Día

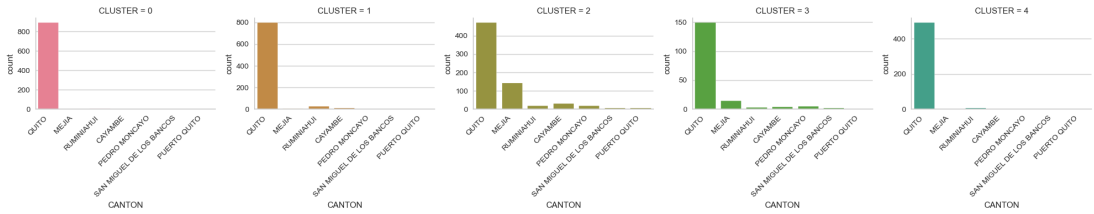
Fuente (Maza. E, 2022)

Figura 7-48 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Periodo



Fuente (Maza. E, 2022)

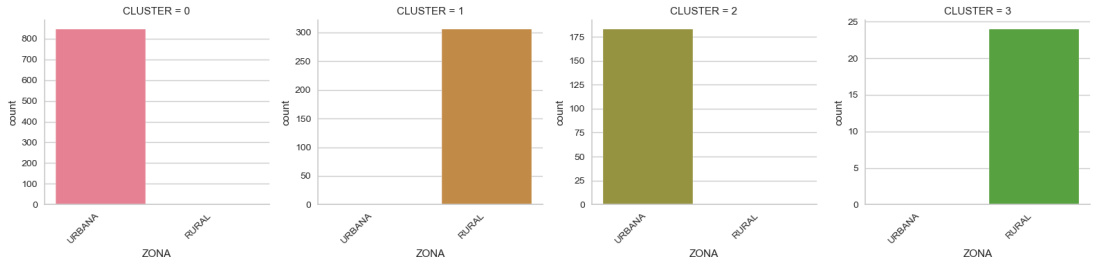
Figura 7-49 Agrupamiento Jerárquico, Pichincha, Clústeres para característica Cantón



Fuente (Maza. E, 2022)

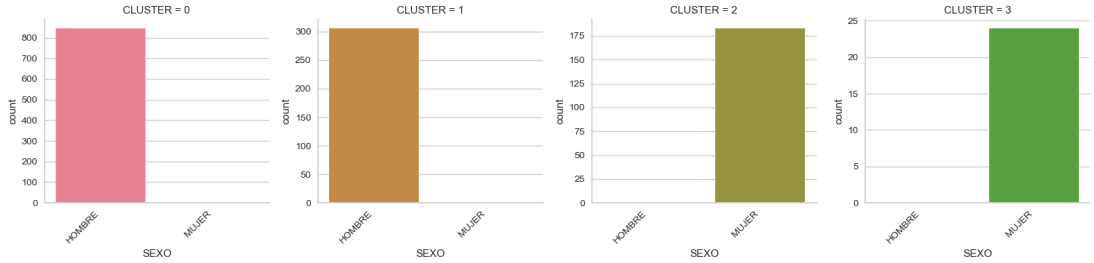
7.1.2.3. Manabí

Figura 7-50 Agrupamiento Jerárquico, Manabí, Clústeres para característica Zona



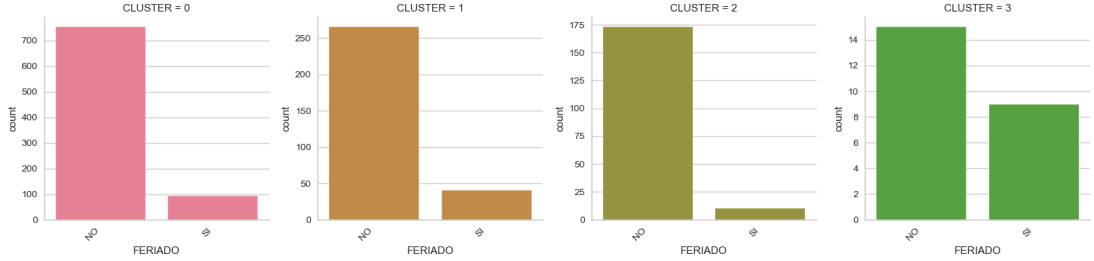
Fuente (Maza. E, 2022)

Figura 7-51 Agrupamiento Jerárquico, Manabí, Clústeres para característica Sexo



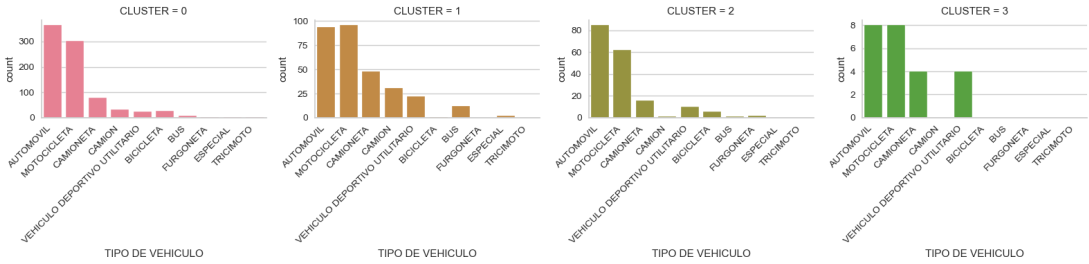
Fuente (Maza. E, 2022)

Figura 7-52 Agrupamiento Jerárquico, Manabí, Clústeres para característica Feriado



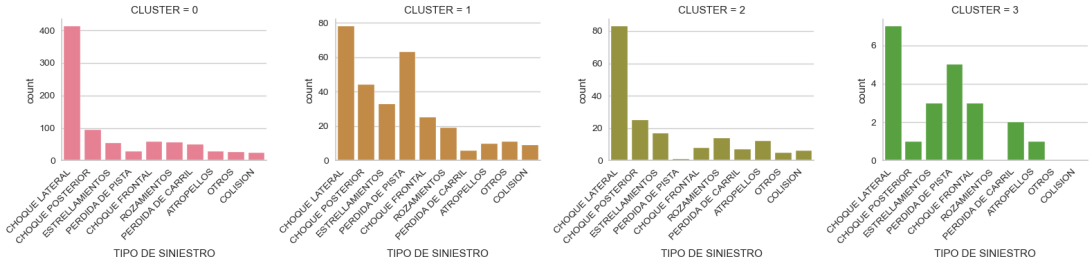
Fuente (Maza. E, 2022)

Figura 7-53 Agrupamiento Jerárquico, Manabí, Clústeres para característica Tipo de Vehículo



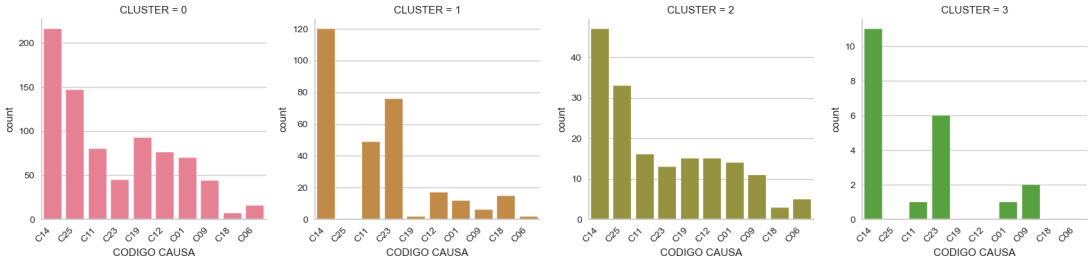
Fuente (Maza. E, 2022)

Figura 7-54 Agrupamiento Jerárquico, Manabí, Clústeres para característica Tipo de Siniestro



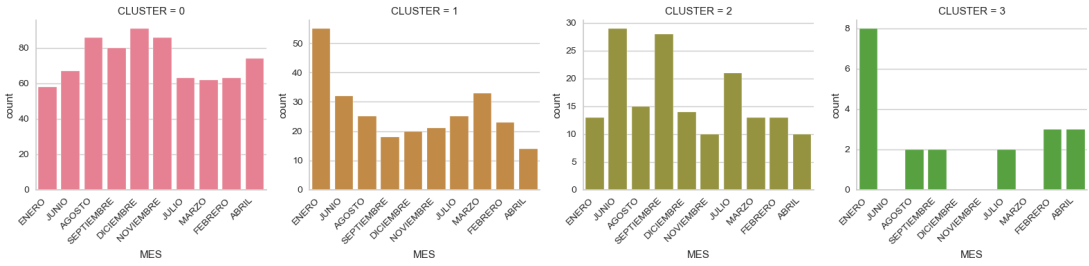
Fuente (Maza. E, 2022)

Figura 7-55 Agrupamiento Jerárquico, Manabí, Clústeres para característica Causa Probable



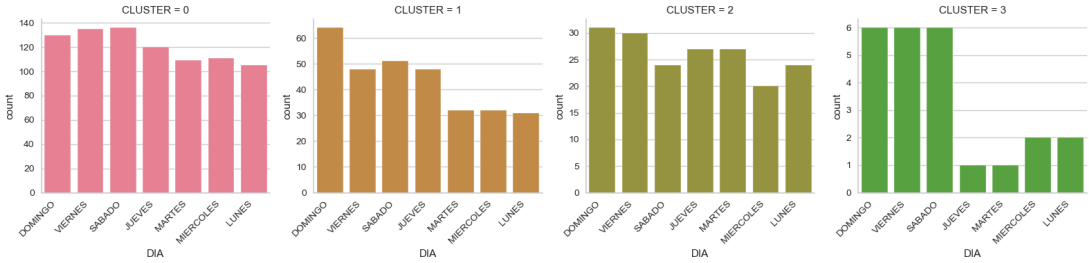
Fuente (Maza. E, 2022)

Figura 7-56 Agrupamiento Jerárquico, Manabí, Clústeres para característica Mes



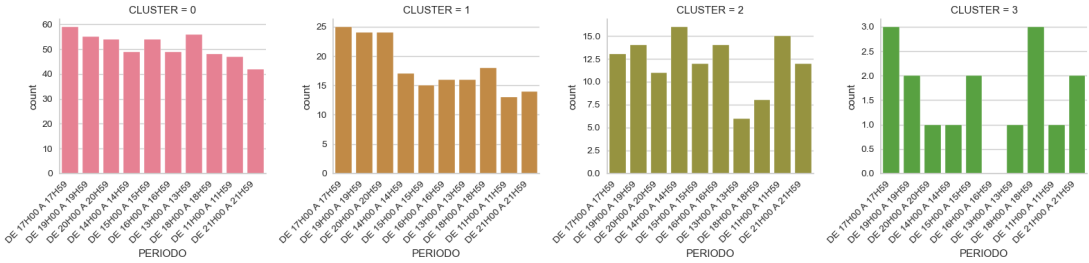
Fuente (Maza. E, 2022)

Figura 7-57 Agrupamiento Jerárquico, Manabí, Clústeres para característica Día



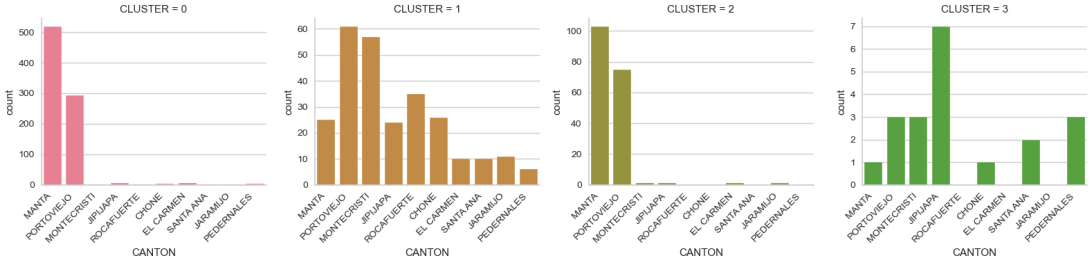
Fuente (Maza. E, 2022)

Figura 7-58 Agrupamiento Jerárquico, Manabí, Clústeres para característica Periodo



Fuente (Maza. E, 2022)

Figura 7-59 Agrupamiento Jerárquico, Manabí, Clústeres para característica Cantón



Fuente (Maza. E, 2022)