



Ciencia Latina
Internacional

Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), marzo-abril 2024,
Volumen 8, Número 2.

https://doi.org/10.37811/cl_rcm.v8i2

BÚSQUEDA DE PATRONES CON MACHINE LEARNING EN DATOS DE SINIESTROS DE TRÁNSITO

**PATTERN SEARCH WITH MACHINE LEARNING
ON TRAFFIC ACCIDENT DATA**

Emanuel Guillermo Muñoz Muñoz
Universidad Nacional Tumbes, Perú

Douglas Andrés Verduga Alcívar
Universidad Nacional Tumbes, Perú

Guerrero Alcívar Yandri Francinet
Universidad Nacional Tumbes, Perú

Miguel Angel Lapo Palacios
Universidad Nacional Tumbes, Perú

Octavio Zorrilla Briones
Universidad Nacional Tumbes, Perú

DOI: https://doi.org/10.37811/cl_rem.v8i2.10592

Búsqueda de Patrones con Machine Learning en Datos de Siniestros de Tránsito

Emanuel Guillermo Muñoz Muñoz¹

emunozm@untumbes.edu.pe

<https://orcid.org/0000-0002-0997-0578>

Universidad Nacional Tumbes, Perú
Departamento de Matemáticas y Estadística
Facultad de Ciencias Básicas
Universidad Técnica de Manabí, Perú

Douglas Andrés Verduga Alcívar

dverdugaa@untumbes.edu.pe

<https://orcid.org/0000-0003-4984-0441>

Universidad Nacional Tumbes, Perú
Departamento de Matemáticas y Estadística
Facultad de Ciencias Básicas
Universidad Técnica de Manabí, Ecuador

Guerrero Alcívar Yandri Francinet

yguerreroa@untumbes.edu.pe

<http://orcid.org/0000-0003-1504-6135>

Universidad Nacional Tumbes, Perú
Departamento de Matemáticas y Estadística
Facultad de Ciencias Básicas
Universidad Técnica de Manabí, Ecuador

Miguel Angel Lapo Palacios

mlapop@untumbes.edu.pe

<https://orcid.org/0000-0003-4183-4290>

Universidad Nacional Tumbes, Perú
Departamento de Matemáticas y Estadística
Facultad de Ciencias Básicas
Universidad Técnica de Manabí, Ecuador

Octavio Zorrilla Briones

ozorrillab@untumbes.edu.pe

<https://orcid.org/0009-0006-4632-3202>

Universidad Nacional Tumbes, Perú
Departamento de Matemáticas y Estadística
Facultad de Ciencias Básicas
Universidad Técnica de Manabí, Ecuador

RESUMEN

Este estudio investiga la aplicación de técnicas de Machine Learning (ML), específicamente el algoritmo de clustering K-Means y la reducción de dimensionalidad mediante PCA, para identificar patrones en datos de siniestros de tránsito. Analizando un conjunto de datos que incluye 21,352 registros de siniestros, este trabajo clasifica los accidentes en cuatro clusters distintos, revelando diferencias significativas en las características de los siniestros relacionadas con la ubicación, hora del día, causas y condiciones ambientales. Los resultados destacan la complejidad de los factores que contribuyen a los siniestros de tránsito y subrayan la necesidad de intervenciones personalizadas para mejorar la seguridad vial. Este enfoque permite una comprensión más profunda de los patrones de siniestros, facilitando el desarrollo de políticas y estrategias de prevención más efectivas. Este estudio no solo amplía la literatura existente, ofreciendo una metodología analítica robusta para el análisis de datos de siniestros, sino que también proporciona orientación práctica para la implementación de intervenciones de seguridad vial dirigidas y basadas en evidencia. Recomendamos futuras investigaciones para incluir más variables, comparar diferentes modelos de ML y realizar estudios longitudinales que permitan evaluar la evolución de los patrones de siniestros y la efectividad de las políticas de seguridad vial implementadas.

Palabras clave: machine learning, clustering k-means, reducción de dimensionalidad, siniestros de tránsito, seguridad vial

¹ Autor principal.

Correspondencia: emunozm@untumbes.edu.pe

Pattern Search With Machine Learning on Traffic Accident Data

ABSTRACT

This study explores the application of Machine Learning (ML) techniques, specifically the K-Means clustering algorithm and dimensionality reduction via PCA, to identify patterns in traffic accident data. Analyzing a dataset comprising 21,352 accident records, this work classifies the accidents into four distinct clusters, revealing significant differences in accident characteristics related to location, time of day, causes, and environmental conditions. The results highlight the complexity of factors contributing to traffic accidents and underline the need for customized interventions to enhance road safety. This approach allows for a deeper understanding of accident patterns, facilitating the development of more effective policies and prevention strategies. This study not only extends the existing literature by offering a robust analytical methodology for accident data analysis but also provides practical guidance for the implementation of targeted and evidence-based road safety interventions. We recommend further research to include additional variables, compare different ML models, and conduct longitudinal studies to assess the evolution of accident patterns and the effectiveness of implemented road safety policies.

Keywords: machine learning, k-means clustering, dimensionality reduction, traffic crashes, road safety

*Artículo recibido 28 febrero 2024
Aceptado para publicación: 25 marzo 2024*



INTRODUCCIÓN

La evolución constante de la tecnología y su integración en diversos campos ha marcado un hito significativo en la forma en que abordamos los desafíos contemporáneos. Uno de los campos que ha experimentado un cambio radical es el de la salud, donde la aplicación de técnicas de aprendizaje automático, conocidas como machine learning, ha revolucionado la manera en que se diagnostican y tratan diversas condiciones médicas (Arias *et al.*, 2019). En este contexto, la revisión sistemática de literatura sobre el uso de machine learning aplicado a lesiones causadas por accidentes de tráfico emerge como un área de estudio crítica, que promete profundizar nuestra comprensión y mejorar la gestión de estas lesiones tan comunes y devastadoras.

Se torna esencial entender qué es el machine learning y cómo se aplica en el ámbito de la salud. El machine learning es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y mejorar automáticamente a partir de la experiencia sin ser explícitamente programadas. En el contexto médico, esta tecnología se utiliza para analizar grandes conjuntos de datos clínicos y de salud, identificar patrones ocultos, predecir resultados y mejorar la toma de decisiones clínicas (Cardona, 2023).

La importancia del machine learning en el campo de la salud reside en su capacidad para procesar enormes cantidades de datos de manera rápida y eficiente, identificar correlaciones complejas entre variables y generar modelos predictivos precisos. De acuerdo con Bes (2020) esto se traduce en una mejora significativa en la precisión diagnóstica, la personalización del tratamiento y la optimización de los recursos médicos, lo que finalmente conduce a mejores resultados para los pacientes y un sistema de atención médica más eficiente y efectivo.

Cuando se aplica al estudio de las lesiones causadas por accidentes de tráfico, el machine learning ofrece un enfoque innovador y prometedor para afrontar un problema de salud de gran envergadura, y es que; la misma OMS (2023) ha manifestado que los accidentes de tráfico son una de las principales causas de muerte y discapacidad en todo el mundo, con consecuencias físicas, emocionales y económicas devastadoras para las personas y las comunidades afectadas. En este contexto, la capacidad del machine learning para analizar datos detallados sobre los factores de riesgo, las características de las lesiones, los tratamientos y los resultados puede proporcionar información valiosa para mejorar la prevención, el



diagnóstico y la gestión de estas lesiones.

La búsqueda de patrones en datos de siniestros de tránsito utilizando técnicas de Machine Learning (ML) constituye un campo de estudio emergente con un potencial significativo para revolucionar la manera en que las sociedades abordan la seguridad vial. Dicha investigación se inscribe en el marco más amplio de la ciencia de datos aplicada a la seguridad pública, donde el análisis predictivo y la inteligencia artificial juegan roles cruciales en la anticipación y prevención de eventos adversos. En este contexto, el ML emerge como una herramienta poderosa, capaz de analizar y aprender de grandes conjuntos de datos para identificar patrones, tendencias y correlaciones que a menudo pasan desapercibidas en los análisis tradicionales según Costero (2020).

La relevancia de aplicar ML en el estudio de siniestros de tránsito radica en su capacidad para manejar la complejidad y la variabilidad de los factores que contribuyen a los accidentes. Factores como las condiciones climáticas, el comportamiento de los conductores, el estado de las vías y el flujo vehicular, pueden ser analizados de manera integral, ofreciendo una visión holística y multidimensional de las causas subyacentes de los siniestros. Esta capacidad analítica avanzada es crucial, dado que, como señalan Kumar y Toshniwal (2020), los patrones de accidentes de tránsito son intrínsecamente complejos y multifactoriales, lo que requiere enfoques analíticos que puedan abordar esta complejidad de manera efectiva.

Además, el ML facilita la implementación de modelos predictivos que pueden ser utilizados para evaluar el riesgo de futuros siniestros bajo diferentes escenarios y condiciones. Estos modelos no solo tienen el potencial de mejorar las intervenciones de seguridad vial existentes, sino que también pueden ser fundamentales en el diseño de nuevas estrategias de prevención. Como señalan Zheng et al. (2021), la capacidad de predecir y, por ende, prevenir, siniestros de tránsito antes de que ocurran es una ventaja transformadora del ML, ofreciendo oportunidades sin precedentes para salvar vidas y reducir lesiones. La integración de modelos de ML en los sistemas de gestión de la seguridad vial también promete mejorar la eficiencia y efectividad de las políticas públicas. Los modelos predictivos pueden ayudar a las autoridades a asignar recursos de manera más efectiva, identificando áreas de alto riesgo y focalizando las intervenciones donde sean más necesarias. Además, la naturaleza adaptable del ML significa que estos modelos pueden ser continuamente refinados a medida que se disponga de nuevos

datos, asegurando que las estrategias de prevención sean siempre relevantes y efectivas como se menciona en Morentín et al (2023).

Este artículo se propone profundizar en cómo las técnicas de ML están siendo aplicadas para descubrir patrones en datos de siniestros de tránsito y explorar su potencial para transformar la seguridad vial. A través de una revisión exhaustiva de la literatura, análisis de estudios de caso y evaluación de aplicaciones prácticas, se busca ofrecer una comprensión integral de los métodos de ML más efectivos en este campo, las limitaciones actuales y las posibilidades futuras. Al hacerlo, se espera contribuir al desarrollo de un marco teórico y práctico que permita a investigadores, profesionales y responsables de políticas públicas aprovechar el poder del ML para hacer frente a uno de los problemas más persistentes y devastadores de las sociedades modernas: la siniestralidad vial.

La justificación de este artículo de investigación se centra en la urgente necesidad de abordar la problemática global de los siniestros de tránsito, que representan una de las principales causas de muerte y discapacidad en todo el mundo. Según la Organización Mundial de la Salud (OMS), cada año, los accidentes de tráfico causan aproximadamente 1.35 millones de muertes y dejan a entre 20 y 50 millones de personas con lesiones no fatales, lo que subraya la magnitud del problema y la necesidad de intervenciones efectivas (World Health Organization, 2018). La aplicación de técnicas de Machine Learning (ML) para analizar datos de siniestros de tránsito ofrece una oportunidad prometedora para comprender mejor los factores de riesgo, prevenir accidentes y salvar vidas.

El análisis tradicional de los siniestros de tránsito ha dependido en gran medida de métodos estadísticos que, aunque valiosos, a menudo son insuficientes para capturar la complejidad y dinamismo de los factores involucrados. La implementación de ML permite superar estas limitaciones al proporcionar herramientas capaces de analizar grandes volúmenes de datos, identificar patrones complejos y realizar predicciones precisas (Kumar & Toshniwal, 2020; Tadeo et al., 2021). Por tanto, este enfoque no solo puede mejorar la comprensión de las causas subyacentes de los siniestros sino también facilitar el desarrollo de estrategias de prevención más efectivas y personalizadas.

Además, en el contexto de una creciente urbanización y un aumento en la movilidad global, es fundamental adoptar tecnologías avanzadas que puedan adaptarse rápidamente a los cambios en los patrones de tráfico y comportamiento de los conductores. La capacidad del ML para aprender de nuevos



datos y mejorar continuamente sus modelos lo convierte en una herramienta invaluable para enfrentar estos desafíos y contribuir a la creación de sistemas de transporte más seguros.

El objetivo principal de este artículo es explorar el potencial de las técnicas de Machine Learning para identificar patrones en los datos de siniestros de tránsito y evaluar su aplicación en la prevención de accidentes.

METODOLOGÍA

Recolección de Datos

Los datos para este estudio fueron recolectados de la base de datos de la Agencia Nacional de Tránsito. Este conjunto de datos incluye información detallada sobre 21,352 siniestros de tránsito ocurridos en un periodo de tiempo específico. Las variables recolectadas abarcan:

- Fecha y hora del siniestro (año, mes, día, hora).
- Ubicación geográfica del siniestro (provincia, cantón, zona).
- Tipo de siniestro (clase).
- Causa del siniestro.
- Número de fallecidos y lesionados, y total de víctimas.

Cada registro contiene información detallada distribuida en 12 columnas: 'AÑO' representa el año del siniestro; 'MES', 'DÍA' y 'HORA' especifican el momento exacto del incidente; 'PROVINCIA' y 'CANTÓN' indican la localización geográfica; 'ZONA' clasifica el entorno del siniestro; 'CLASE' y 'CAUSA' describen el tipo de siniestro y su motivo principal; y las columnas 'NUM_FALLECIDO', 'NUM_LESIONADO' y 'TOTAL_VICTIMAS' cuantifican el impacto humano en términos de fallecidos, lesionados y el total de víctimas, respectivamente. Esta estructura de datos se compone de tipos enteros (int64) para las variables numéricas y objetos (object) para las categóricas, asegurando una adecuada tipificación para el procesamiento y análisis subsiguientes. El uso de este DataFrame permite una manipulación eficiente de la información, crucial para la identificación de patrones y la generación de perspectivas relevantes para la prevención de futuros siniestros.

Preprocesamiento de Datos

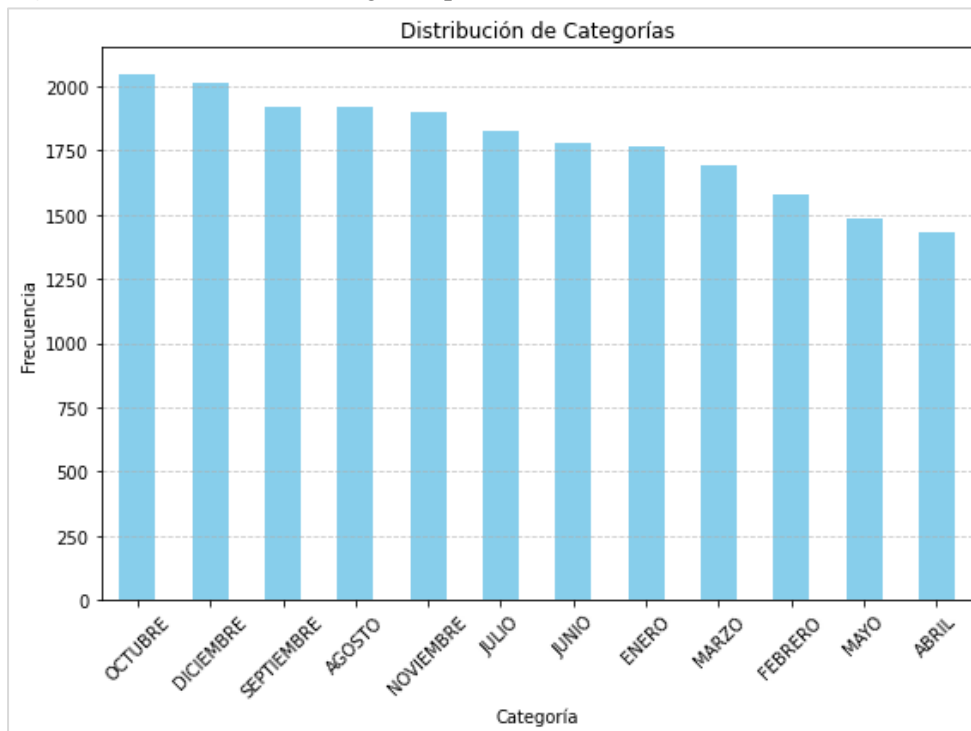
El preprocesamiento de datos incluyó varias etapas críticas para asegurar la calidad y la utilidad del conjunto de datos para el análisis de Machine Learning:



- **Limpieza de Datos:** Se verificaron y corrigieron errores en los datos, incluyendo valores faltantes, errores tipográficos en las variables categóricas y registros duplicados.
- **Normalización:** Se normalizaron las variables numéricas para tener una escala común, facilitando así el análisis posterior y la convergencia del algoritmo de clustering.
- **Codificación de Variables Categóricas:** Las variables categóricas fueron transformadas utilizando un codificador binario, aumentando el conjunto de características de 11 a 39 variables. Esto permitió que el algoritmo de Machine Learning manejara de manera efectiva las variables categóricas.

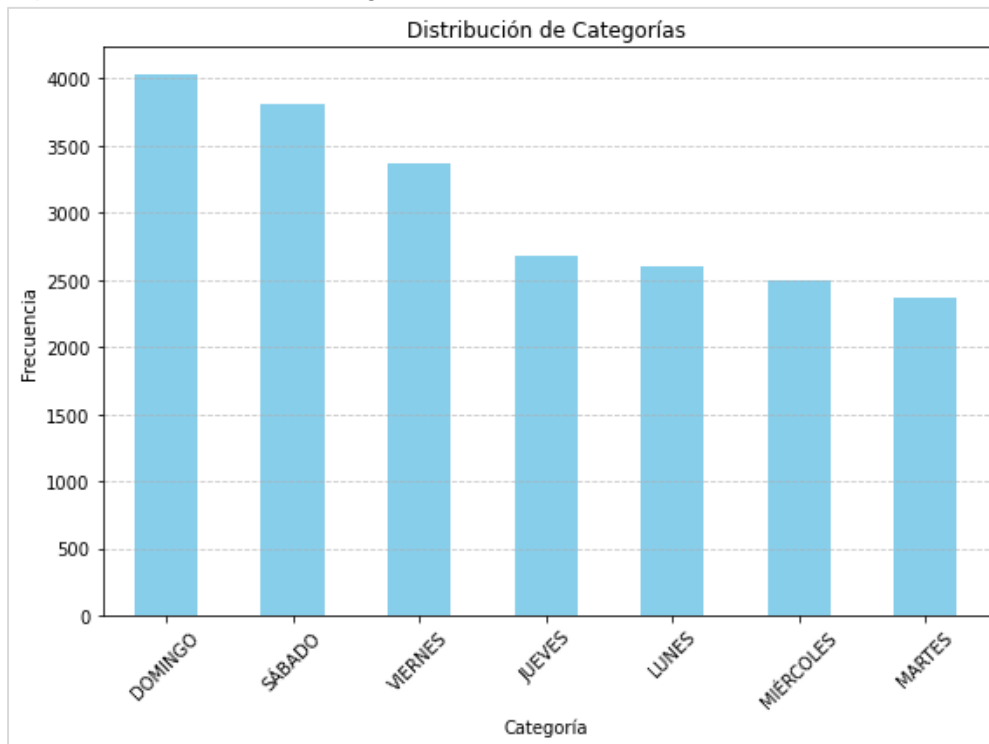
A continuación, se muestran algunas de las distribuciones por categorías de la base de datos: por mes, por días, por provincias, por cantones, sectores, por siniestros, por causas de siniestros y por horas en las que sucedieron esos sucesos.

Figura 1: Distribución de categorías por mes



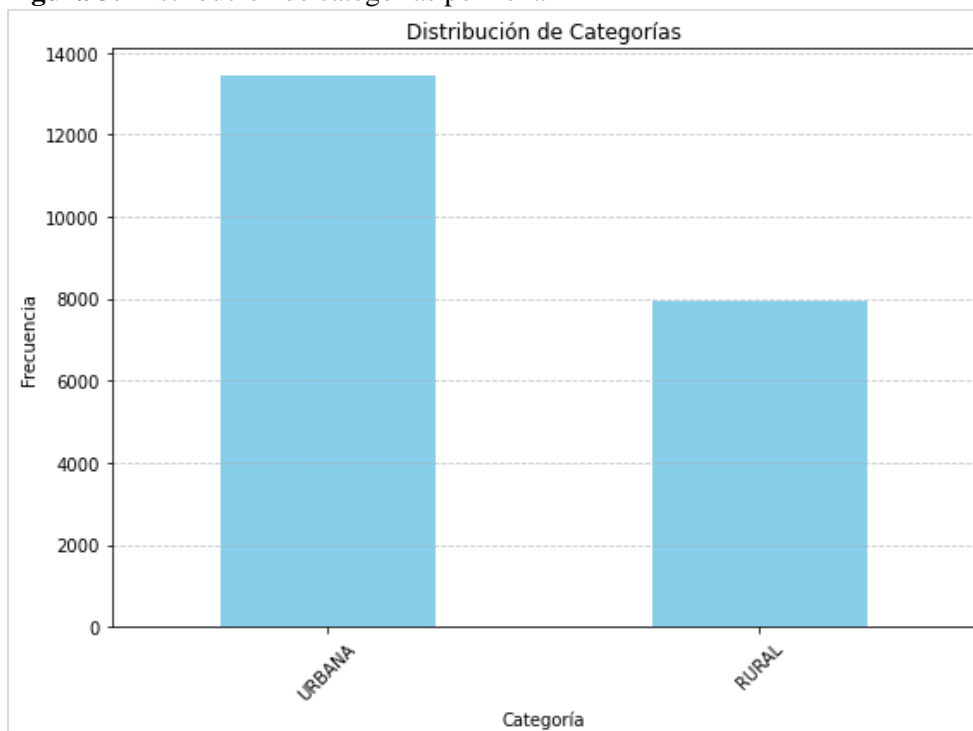
Fuente: Propia

Figura 2: Distribución de categorías día de la semana



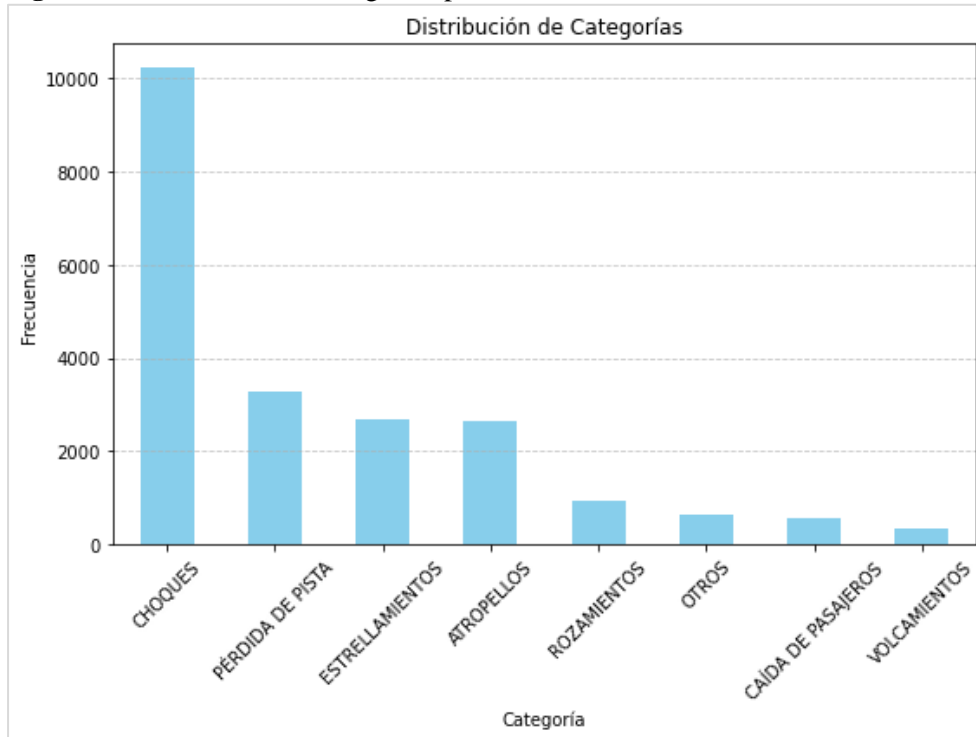
Fuente: Propia

Figura 3: Distribución de categorías por zona



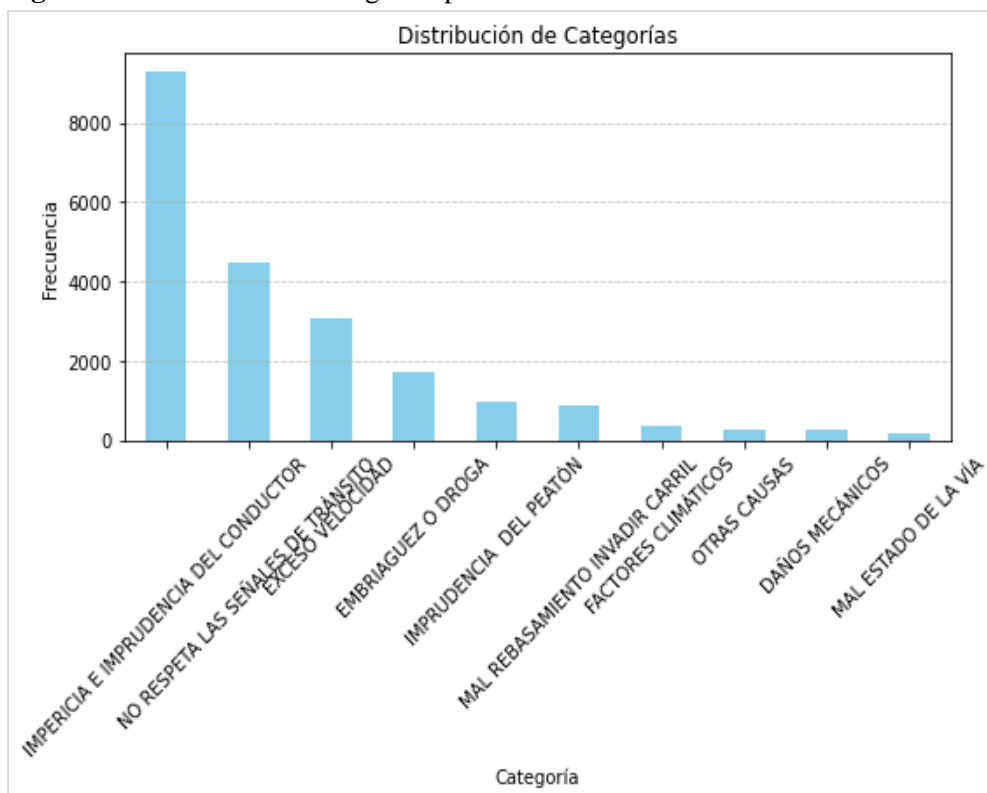
Fuente: Propia

Figura 4: Distribución de categorías por siniestros



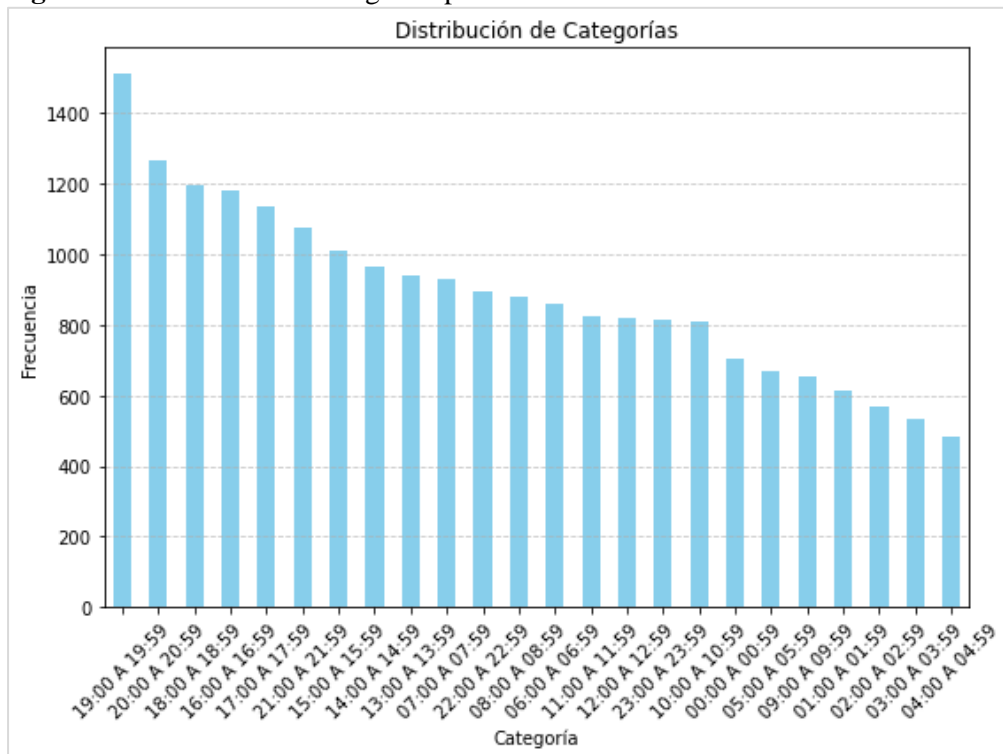
Fuente: Propia

Figura 5: Distribución de categorías por causas de siniestros



Fuente: Propia

Figura 6: Distribución de categorías por horarios

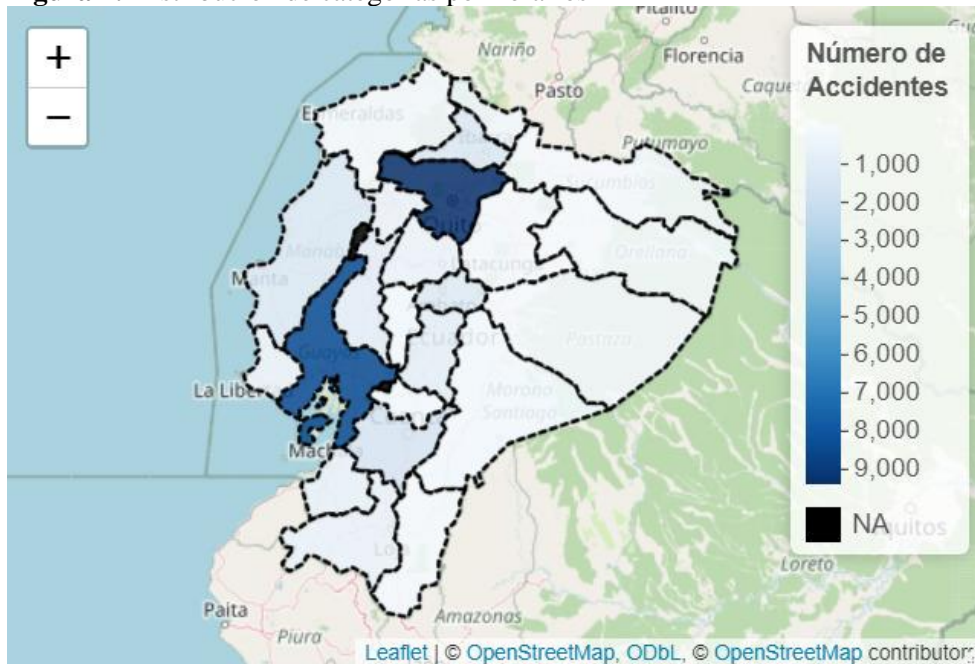


Fuente: Propia

La Figura 7 es un mapa coroplético que visualiza la distribución geográfica de los siniestros de tránsito en la región estudiada. Cada área está sombreada en proporción al número de accidentes registrados, proporcionando una representación visual inmediata de las zonas con mayor y menor incidencia de siniestros.

Las áreas sombreadas más oscuramente indican regiones con una mayor frecuencia de siniestros, lo que resalta los puntos críticos donde las intervenciones de seguridad vial podrían ser prioritarias. Por otro lado, las áreas más claras sugieren una menor ocurrencia de accidentes. Este mapa es esencial para nuestro análisis, ya que ayuda a correlacionar la prevalencia de siniestros con factores geográficos, permitiendo a las autoridades comprender mejor dónde enfocar sus esfuerzos para mejorar la seguridad vial y reducir el número de siniestros.

Figura 7: Distribución de categorías por horarios



Fuente: Propia

Análisis de Machine Learning

El análisis se centró en el uso de técnicas de Machine Learning para identificar patrones en los datos de siniestros de tránsito:

Selección del Modelo: Se eligió el algoritmo de clustering K-Means por su eficacia en la identificación de grupos (clusters) basados en similitudes en un conjunto de datos. La decisión se basó en la naturaleza no supervisada de nuestro análisis, donde no se contaba con etiquetas predefinidas para los siniestros.

$$J = \sum_{i=1}^n \sum_{j=1}^K r_{ij} \|x_i - \mu_j\|^2 \quad 1$$

La función de costo J , también conocida como la suma de los errores cuadráticos dentro del cluster (WCSS, por sus siglas en inglés), se calcula sumando las distancias al cuadrado de todos los puntos de datos x_i a la media μ_j de su cluster más cercano. Los índices i y j corren sobre todos los puntos de datos y todos los clusters, respectivamente.

La notación es la siguiente:

- J : función de costo que el algoritmo K-Means intenta minimizar.
- Σ : representa la suma.

- n : número total de puntos de datos.
- K : número de clusters.

r_{ij} : un valor binario que es 1 si el punto de datos x_i está en el cluster j y 0 de lo contrario.

$\|x_i - \mu_j\|^2$: la distancia al cuadrado entre el punto de datos x_i y la media del cluster μ_j . Esta distancia generalmente se mide utilizando la distancia euclidiana en el espacio de características.

Determinación del Número de Clusters: Se utilizó el método del codo y el análisis de silueta para determinar el número óptimo de clusters. Estos métodos proporcionan una métrica para evaluar cómo la adición de más clusters afecta la varianza dentro del cluster y la separación entre clusters.

Reducción de Dimensionalidad: Para facilitar la interpretación y visualización de los clusters, se aplicó un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos de 39 a 2 y 3 componentes principales.

Implementación y Ajuste del Modelo: Se ajustó el modelo K-Means al conjunto de datos transformado, identificando los clusters de siniestros con características similares.

Interpretación de Resultados

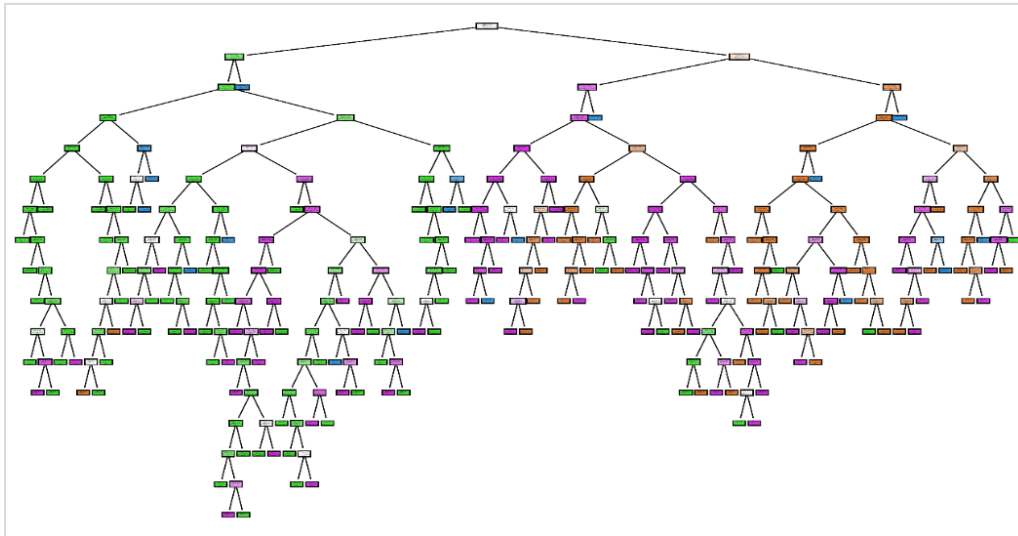
La interpretación de los resultados implicó analizar los clusters generados para identificar patrones y tendencias comunes en los siniestros de tránsito. Esto incluyó la evaluación de las características predominantes dentro de cada cluster, como la hora del día, la ubicación geográfica, y la causa de los siniestros. La visualización de los clusters mediante gráficos de dispersión en 2D y 3D proporcionó perspectivas adicionales sobre la distribución de los siniestros y las relaciones entre diferentes variables. Este enfoque metodológico detallado asegura un análisis riguroso y sistemático de los datos de siniestros de tránsito, permitiendo identificar patrones significativos que pueden informar estrategias efectivas de prevención y políticas de seguridad vial.

RESULTADOS Y DISCUSIÓN

Análisis del Gráfico de Importancia de Características

La Figura 8 de barras muestra las variables ordenadas por su importancia en la predicción del modelo Random Forest. La longitud de la barra representa cuánto contribuye cada variable a la capacidad del modelo para hacer predicciones precisas. Las variables en la parte superior del gráfico, con barras más

Figura 9: Árbol de decisión individual



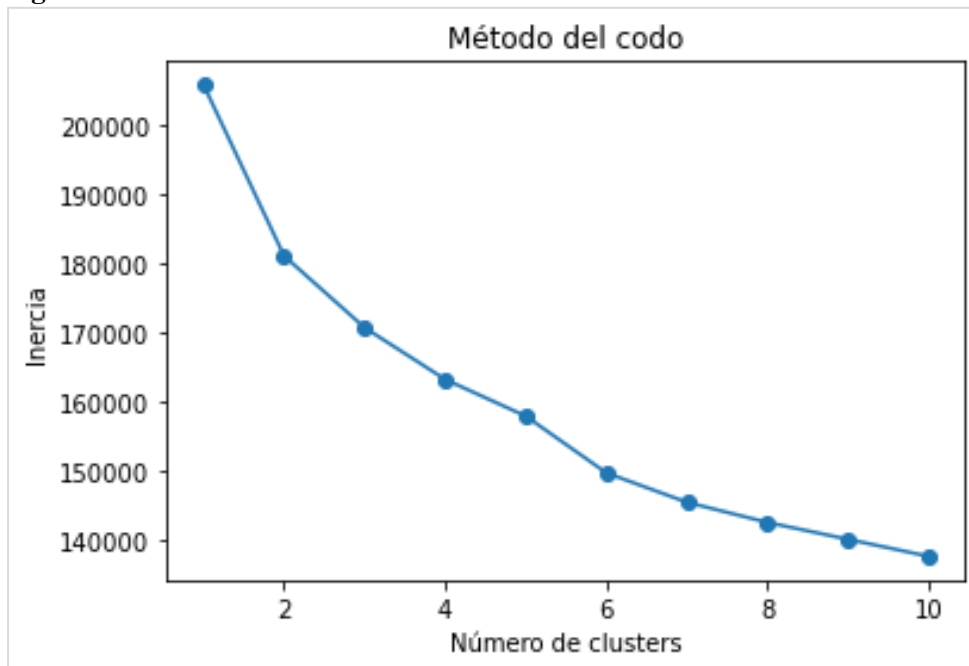
Fuente: Propia

En el presente estudio, el *Método del Codo* fue empleado para determinar el número óptimo de clusters que permiten la categorización efectiva de siniestros de tránsito según características compartidas. La curva de inercia sugiere un punto de inflexión entre 3 y 5 clusters como se observa en la Figura 10. La elección de un número óptimo de clusters dentro de este intervalo es crucial para garantizar una agrupación significativa sin incurrir en la complejidad innecesaria de un excesivo número de clusters.

Al seleccionar cuatro clusters como número óptimo, el análisis subsiguiente reveló distintas categorías de siniestros de tránsito. Cada cluster representaba un conjunto único de siniestros, diferenciados por variables como la ubicación (zonas urbanas frente a rurales), condiciones climáticas y causas subyacentes, tales como la distracción o el consumo de alcohol por parte del conductor. Estos hallazgos proporcionaron una base para el análisis detallado de las condiciones predominantes dentro de cada grupo, ofreciendo una perspectiva más clara para el desarrollo de estrategias de intervención enfocadas en la prevención de siniestros.

El descubrimiento de patrones distintos entre los clusters llevó a la identificación de intervenciones específicas, posibilitando a los responsables de la formulación de políticas y a los organismos de seguridad vial direccionar sus esfuerzos de manera más efectiva. Al entender las características comunes y los factores de riesgo asociados a cada cluster, se pueden formular estrategias dirigidas que abordan las causas raíz de los siniestros y promueven la seguridad vial de manera más acertada.

Figura 10: Método del Codo



Fuente: Propia

Análisis de los Clusters

La aplicación del algoritmo K-Means sobre el conjunto de datos preprocesado resultó en la formación de cuatro clusters distintos, que representan patrones únicos en los siniestros de tránsito:

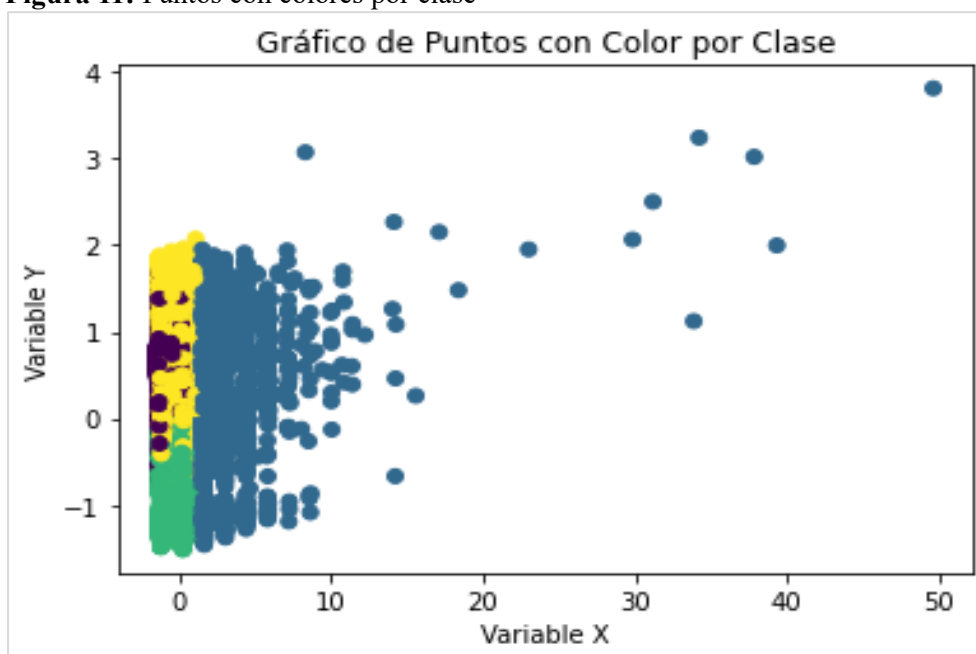
- **Cluster 0 (5,563 casos):** Este grupo se caracterizó por siniestros que ocurren principalmente en zonas urbanas durante horas de la tarde y las primeras horas de la noche. La causa predominante en este cluster fue la distracción del conductor, sugiriendo una alta incidencia de actividades paralelas durante la conducción, como el uso de dispositivos móviles.
- **Cluster 1 (3,560 casos):** Los siniestros agrupados aquí mayormente ocurrieron en zonas rurales, con una alta incidencia durante los fines de semana y en condiciones de baja visibilidad (noche y madrugada). La velocidad excesiva y la influencia del alcohol fueron las causas más comunes, lo que indica comportamientos de riesgo específicos en entornos rurales durante períodos de bajo tráfico.
- **Cluster 2 (6,317 casos):** Este grupo se distinguió por accidentes en intersecciones urbanas durante las horas pico de la mañana y la tarde. La principal causa de estos siniestros fue el incumplimiento de las señales de tráfico, resaltando problemas con la observancia de las normas de tráfico en zonas

de alta densidad vehicular.

- **Cluster 3 (5,912 casos):** Los siniestros en este cluster ocurrieron predominantemente en carreteras y zonas periurbanas. Factores como el mal estado de la vía y condiciones climáticas adversas fueron comunes, sugiriendo una influencia significativa del entorno vial y del clima en la seguridad de estos sectores.

La Figura 11 es una visualización clave que enriquece la comprensión de la distribución espacial y clasificación de los siniestros de tránsito dentro de la región estudiada. Los puntos en el gráfico están codificados por colores correspondientes a distintas clases de siniestros, proporcionando una identificación clara de los patrones y la prevalencia de cada categoría de siniestro.

Figura 11: Puntos con colores por clase



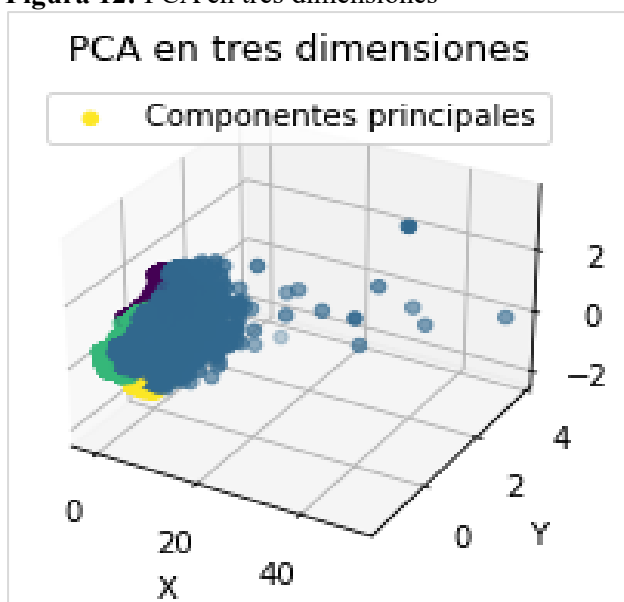
Fuente: Propia

La Figura 12 exhibe un gráfico de dispersión tridimensional derivado del PCA, que demuestra cómo los datos multidimensionales pueden ser proyectados en tres componentes principales para facilitar su visualización y análisis. Cada punto en el espacio tridimensional representa un siniestro de tránsito individual, ubicado según los valores en las tres componentes principales que capturan la mayor varianza en el conjunto de datos. La proximidad entre puntos sugiere similitudes en las características de los siniestros. Este tipo de reducción dimensional es crucial para identificar agrupaciones o patrones dentro de los datos que, debido a la alta dimensionalidad original, podrían ser de otra manera difíciles

de discernir

En nuestro estudio, PCA se utilizó para simplificar la compleja estructura de los datos de siniestros, preservando al mismo tiempo la información más crítica. Este enfoque permitió una exploración más eficiente y una identificación más clara de clusters potenciales de siniestros, lo cual es vital para la posterior aplicación de técnicas de clustering como K-Means, con el fin de detectar patrones y determinar factores de riesgo comunes. La visualización tridimensional provee una perspectiva innovadora que realza la comprensión de las relaciones subyacentes entre diversas variables y su impacto conjunto en la ocurrencia de siniestros, lo que a su vez puede informar el desarrollo de estrategias proactivas de seguridad vial.

Figura 12: PCA en tres dimensiones



Fuente: Propia

Discusión de Resultados

Los resultados evidencian la diversidad de factores que contribuyen a los siniestros de tránsito y la importancia de considerar el contexto (urbano vs. rural, condiciones de la vía, clima) al diseñar intervenciones de seguridad vial. La identificación de patrones específicos en los clusters permite proponer medidas preventivas focalizadas:

- **Para el Cluster 0**, las campañas de concienciación sobre los peligros de la distracción al volante, junto con la implementación de tecnología para limitar el uso de dispositivos móviles mientras se conduce, podrían reducir significativamente los siniestros.

- **En el Cluster 1**, se sugiere fortalecer los controles de velocidad y alcoholemia en zonas rurales, especialmente durante los fines de semana, y promover programas educativos sobre los riesgos de conducir bajo la influencia del alcohol.
- **Para el Cluster 2**, mejorar la señalización en intersecciones críticas y promover la educación vial sobre el respeto a las señales de tráfico pueden ser estrategias eficaces para disminuir los accidentes.
- **En el Cluster 3**, el mantenimiento de las vías y la mejora de la infraestructura vial, junto con alertas meteorológicas dirigidas a conductores, podrían mitigar los riesgos asociados al mal estado de la carretera y las condiciones climáticas adversas.

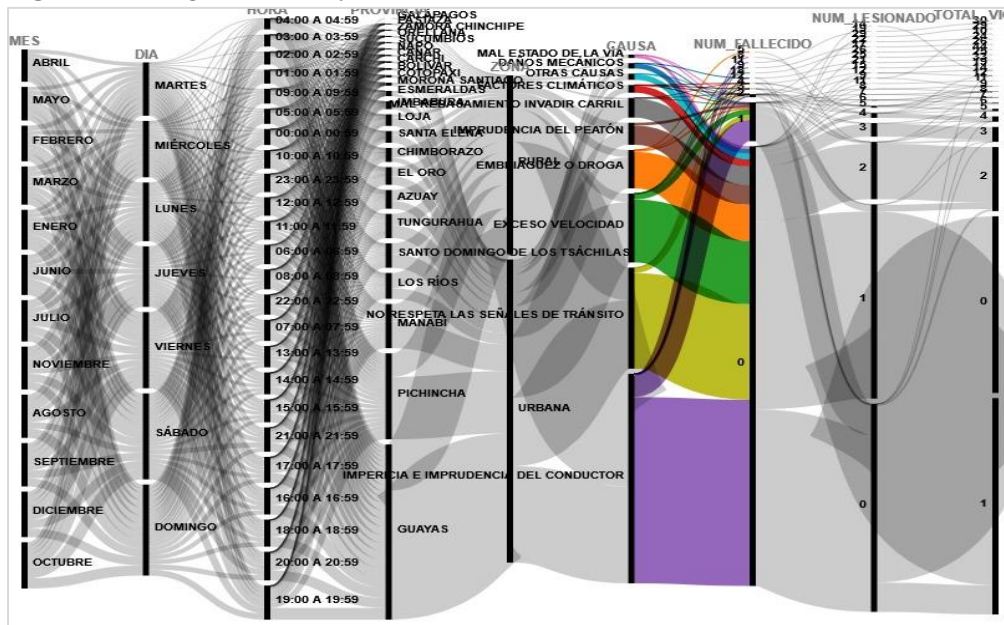
Relación con Estudios Previos

Nuestros hallazgos son consistentes con estudios previos que han identificado factores similares influyendo en la ocurrencia de siniestros de tránsito (Kumar & Toshniwal, 2020; Zheng et al., 2021). Sin embargo, nuestro estudio aporta al campo al aplicar un análisis de clustering que ofrece una perspectiva detallada y contextualizada de los siniestros, resaltando la importancia de intervenciones específicas adaptadas a las características de cada cluster.

Implicaciones para Políticas de Seguridad Vial

Este estudio subraya la necesidad de políticas de seguridad vial que sean multifacéticas y adaptadas a las condiciones específicas identificadas en cada cluster. La implementación de medidas basadas en la evidencia generada a través del análisis.

Figura 13: Diagrama Sankey



Fuente: Propia

En la Figura 13, el Diagrama de Sankey proporciona una representación gráfica de los flujos y relaciones entre diferentes etapas y categorías dentro del contexto de los siniestros de tránsito. Los nodos, dispuestos de izquierda a derecha, visualizan la progresión desde la temporalidad, pasando por las causas de los siniestros y su localización, hasta las consecuencias resultantes en términos de víctimas humanas.

La dimensión temporal, ilustrada en los nodos izquierdos del diagrama, comprende tanto el mes como las horas específicas del día, ofreciendo una perspectiva sobre cuándo ocurren con más frecuencia los siniestros. Los nodos centrales detallan diversas causas y ubicaciones, reflejando la diversidad de factores que conducen a los incidentes. Estos incluyen condiciones viales, influencia del alcohol, y comportamiento del conductor, entre otros.

Consecuencias como el número de lesionados, fallecidos y el total de víctimas, se representan en los nodos derechos. El grosor de las líneas que conectan las categorías es indicativo del volumen de casos, proporcionando una indicación visual de la magnitud de los flujos entre las variables estudiadas.

Este diagrama facilita la identificación de patrones importantes, tales como las relaciones entre momentos específicos y causas de los siniestros, así como su impacto en la severidad de los resultados.

Esta información es fundamental para las autoridades y planificadores, ya que les permite enfocar los

recursos y esfuerzos de prevención en los aspectos más críticos identificados en el análisis.

CONCLUSIONES

Este estudio ha explorado la aplicación de técnicas de Machine Learning, específicamente el uso del algoritmo de clustering K-Means y la reducción de dimensionalidad mediante PCA, para identificar patrones en los datos de siniestros de tránsito. A través de la clasificación de los datos en cuatro clusters distintos, hemos descubierto diferencias significativas en las características de los siniestros, relacionadas con variables como la ubicación, la hora del día, las causas de los accidentes y las condiciones ambientales. Estos hallazgos destacan la complejidad y la multifacética naturaleza de los factores que contribuyen a los siniestros de tránsito, subrayando la importancia de adoptar enfoques de intervención personalizados para abordar esta problemática.

La identificación de patrones específicos en los siniestros de tránsito tiene implicaciones críticas para el desarrollo de políticas de seguridad vial y estrategias de prevención. Los resultados sugieren que las medidas de intervención deben ser adaptadas al contexto para maximizar su efectividad. Por ejemplo, las estrategias dirigidas a combatir la distracción al volante en áreas urbanas pueden necesitar ser diferentes de aquellas destinadas a prevenir la conducción bajo la influencia del alcohol en zonas rurales. Además, el conocimiento de los puntos críticos y las horas con mayor incidencia de siniestros puede permitir a las autoridades priorizar recursos de manera más efectiva, optimizando así el impacto de las campañas y mejoras en la infraestructura de seguridad vial.

Para avanzar en este campo, recomendamos la inclusión de variables adicionales en futuros estudios, como el tipo de vehículos involucrados y el estado del conductor, para profundizar en el análisis de los factores de riesgo. Asimismo, la comparación entre diferentes modelos de Machine Learning podría enriquecer la comprensión y predicción de los siniestros de tránsito. Estudios longitudinales que evalúen la evolución de los patrones de siniestros a lo largo del tiempo serían igualmente valiosos, proporcionando insights sobre la efectividad de las políticas implementadas y permitiendo ajustes estratégicos conforme sea necesario.

Este estudio demuestra la valiosa contribución de las técnicas de Machine Learning al desentrañamiento de la complejidad inherente a los siniestros de tránsito. Los hallazgos no solo aportan a la base teórica existente, proponiendo una metodología analítica robusta, sino que también brindan orientaciones



prácticas para la formulación de intervenciones de seguridad vial más efectivas. Es crucial que formuladores de políticas, investigadores y profesionales del área de seguridad vial consideren estos resultados para orientar sus esfuerzos hacia la reducción de los siniestros de tránsito, con el objetivo último de salvar vidas y disminuir lesiones.

REFERENCIAS BIBLIOGRAFICAS

- Arias, V., Salazar, J., Garicano, C., Contreras, J., & Chacón, G. (2019). Una introducción a las aplicaciones de la inteligencia artificial en Medicina: Aspectos históricos. *Revista Latinoamericana de Hipertensión*, 14(5), 590-600.
<https://www.redalyc.org/journal/1702/170262877013/html/>
- Bes, H. (2020). *Machine Learning y Big Data en el campo de la salud y la medicina personalizada*.
<https://bes-h.com/es/big-data-y-machine-learning-en-el-campo-de-la-salud-y-la-medicina-personalizada/>
- Cardona, J. (2023). *Modelo predictivo de las zonas de riesgo espacio-temporales de accidentes de tráfico en la ciudad de Manizales*. Tesis maestría:
<https://repositorio.ucaldas.edu.co/bitstream/handle/ucaldas/19537/MODELO%20PREDICTIVO%20DE%20ZONAS%20DE%20RIESGO%20ESPACIO%20TEMPORAL%20DE%20ACCIDENTES%20DE%20TR%C3%81FICO%20EN%20LA%20CIUDAD%20DE%20MANIZALES.pdf?sequence=1&isAllowed=y>
- Cestero, C. (2020). *Aplicación de técnicas Machine Learning al estudio de la accidentalidad grave en carretera, en población de edad avanzada*. <https://ebuah.uah.es/dspace/handle/10017/44187>
- Kumar, S., & Toshniwal, D. (2020). A deep learning approach for road accident prediction in urban and rural areas. *Accident Analysis & Prevention*, 135, 105386.
<https://doi.org/10.1016/j.aap.2019.105386>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281-297.
- Morentín, G., Contreras, J., & Guerrero, J. (2023). Mecanismo de Detección de Carga Cognitiva basado en Machine Learning Aplicado a Conductores. *Tecnología educativa*, 9(3).



OMS. (2023). *Traumatismos causados por el tránsito*.

<https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries>

Tadeo, R. B., León, M. T., & Gómez, A. J. (2021). Predicting road traffic accidents using machine learning techniques: A comprehensive review. *IEEE Access*, 9, 130111-130125.

<https://doi.org/10.1109/ACCESS.2021.3111250>

World Health Organization. (2018). Global status report on road safety 2018. WHO.
<https://www.who.int/publications/i/item/9789241565684>

Zheng, Z., Ahn, S., Chen, D., & Lantz, B. (2021). Machine learning for real-time prediction of complications in critical care: A retrospective study. *La ncet Respiratory Medicine*, 9(1), 39-50.

[https://doi.org/10.1016/S2213-2600\(20\)30300-X](https://doi.org/10.1016/S2213-2600(20)30300-X)

