PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN SISTEMA DE INFORMACIÓN, MENCIÓN DATA SCIENCE

TESIS

"Implementación de un Algoritmo ML de Aprendizaje No Supervisado para la creación de Un Cronograma de Mantenimiento Vial de las vías Rurales de la Provincia de Santo Domingo de los Tsáchilas".

AUTOR: Martín Mauricio Pérez Huachamboza, Ing.

DIRECTOR: Eduardo José Montero Bermúdez, Ing., MSc.

Quito - 2023

DEDICATORIA

A mi familia, por su amor incondicional y apoyo constante en cada paso de mi camino. A mis profesores y mentores, por iluminar mi sendero con su sabiduría y paciencia. A mis amigos, por estar siempre ahí, en los momentos de alegría y desafío. Dedico este trabajo a todos aquellos que creyeron en mí cuando yo mismo dudé y a aquellos que me enseñaron el verdadero valor de la perseverancia y la dedicación. Que este logro sea un reflejo de nuestra unión y esfuerzo compartido.

AGRADECIMIENTO

Quisiera expresar mi más profundo agradecimiento a todas las personas que han hecho posible este proyecto. Primero, a mi familia, cuyo amor y apoyo han sido el faro en este viaje. A mis profesores y asesores, por su invaluable orientación y conocimiento. A mis compañeros, por el espíritu de colaboración y amistad. Y a todas las personas que, de una forma u otra, han contribuido a mi crecimiento personal y profesional. Este logro es también suyo, y les estoy eternamente agradecido.

Con gratitud,

Martín Pérez H.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Resumen

Este trabajo investiga la optimización del mantenimiento vial en Santo Domingo de los Tsáchilas mediante el uso de algoritmos de *Machine Learning* no supervisados, específicamente a través del análisis de clústeres. Utilizando la metodología CRISP-DM, se analizaron datos históricos y actuales de las vías, enfocándose en variables como longitud, estado de la vía, y producción económica, para identificar patrones y priorizar intervenciones de mantenimiento. Los resultados muestran una segmentación efectiva de las vías que permite una planificación de mantenimiento más precisa y enfocada.

Palabras clave: Optimización del mantenimiento vial, *Machine Learning* no supervisados, Análisis de clústeres, Metodología CRISP-DM, Planificación de mantenimiento

Abstract

This study explores the optimization of road maintenance in Santo Domingo de los Tsáchilas using unsupervised Machine Learning algorithms, focusing on clustering analysis. Applying CRISP-DM methodology, historical and current data of the roads were analyzed, emphasizing on variables such as length, road condition, and economic production, to identify patterns and prioritize maintenance interventions. The findings demonstrate an effective segmentation of the roads, bringing more precise and targeted maintenance planning.

Keywords: Road maintenance optimization, Unsupervised Machine Learning, Clustering analysis, CRISP-DM methodology, Maintenance planning

Capítulo 1

Introducción

La Provincia de Santo Domingo de los Tsáchilas, aunque jurídicamente establecida hace 14 años, lleva más de 132 años forjando su identidad territorial. En este largo periodo ha habido un desarrollo acelerado y a menudo no planificado, especialmente en lo que respecta al uso del suelo y a la infraestructura vial. Esta historia ha dado forma a una red de caminos con características únicas y desafíos específicos.

Actualmente, la comunidad enfrenta retos significativos relacionados con su sistema vial. La responsabilidad de gestionar y mantener la red vial rural, que incluye una diversidad de superficies como asfalto, tierra y lastre, recae en el Gobierno Provincial de los cantones de Santo Domingo y La Concordia. El lastre, en particular, predomina en gran parte de la red, presentando desafíos únicos de mantenimiento.

Durante las temporadas de lluvia, los caminos rurales experimentan un rápido deterioro, lo que pone a prueba la capacidad del Gobierno Provincial para su mantenimiento y atención oportuna. Esta situación subraya la necesidad de un monitoreo constante y efectivo del estado de las vías, especialmente en la preparación para la temporada invernal. Un enfoque proactivo en este aspecto es vital para asegurar la prestación de servicios adecuados a la comunidad y garantizar la seguridad y eficiencia de la red vial en estos períodos críticos.

En este contexto, se hace evidente la importancia de una planificación y gestión vial basada en datos precisos y análisis detallados, para abordar de manera efectiva necesidades específicas de la infraestructura vial de Santo Domingo de los Tsáchilas.

1.1 Antecedentes.

El sistema vial provincial rural tiene un total de 2.953,86 Km, de los cuales el 82.6% es lastrado; 6.1% es tierra y el 11% es pavimento (Inventario Vial 2021). Estas cifras muestran un déficit de intervención del mejoramiento de vías debido que el 82.6%; de las vías de la provincia no tienen un tratamiento definitivo de sus capas de rodadura en las diferentes parroquias (PDOT – SDT, 2020, pág. 456).

Tabla 1. Red Vial Provincial de Santo Domingo de los Tsáchilas

Tipo de Rodadura	Longitud (m)	Longitud (km)	Porcentaje
Asfalto	325.000,0	325,0	11,0%
Lastre	2.439.390,0	2.439,4	82,6%
Tierra	181.140,0	181,1	6,1%
Otros	8.330,0	8,3	0,3%
Total	**2.953.860,0**	**2.953,9**	**100,0%**

Fuente: Plan de Desarrollo y Ordenamiento Territorial (PDOT) 2020 - 2030

Por otro lado, aunque la provincia se divide en 10 parroquias, existen zonas de planificación y áreas de cantones que son consideradas rurales. En las cifras expuestas es evidente que existen más vías de tipo lastre con más del 80%, y por consecuencia son las vías que más requieren de servicios de mantenimiento.

Tabla 2. Superficie de capa de rodadura por parroquia en km

Parroquia	Asfalto	Lastre	Tierra	Otros	Total (km)
MONTERREY	2,2	159,7	4,0	0,3	166,2
LA VILLEGAS	4,4	44,8	1,6	0,1	50,9
PLAN PILOTO	11,3	54,2	5,3	0,1	70,9
LA CONCORDIA	-	49,7		0,1	49,8
VALLE HERMOSO	27,9	225,9	45,3	0,6	299,7
SAN JACINTO DEL BÚA	22,6	228,7	26,4	1,3	278,9
SANTA MARÍA DEL TOACHI	25,9	161,6	15,8	0,5	203,7
LUZ DE AMÉRICA	23,0	132,0	3,0	0,4	158,3
PUERTO LIMÓN	40,4	234,6	18,5	0,5	293,9
EL ESFUERZO	30,7	153,4	48,5	0,7	233,3
ALLURIQUIN	23,4	443,2	3,3	0,9	470,7
ZONAS DE PLANIFICACION	113,2	551,8	9,5	3,1	677,5
Total	**325**	**2439,39**	**181,14**	**8,33**	**2953,86**

Fuente: Plan de Desarrollo y Ordenamiento Territorial (PDOT) 2020 - 2030

1.2 Justificación

En el escenario empresarial contemporáneo, la innovación continua en procesos y sistemas no es meramente un ideal, sino una necesidad preeminente para mantener una ventaja competitiva en el mercado. Las organizaciones están persistentemente en la búsqueda de tecnologías emergentes que puedan brindar soluciones eficaces y eficientes. En este contexto, el *Machine Learning* (ML) se ha establecido como una herramienta formidable para el análisis y la predicción en grandes volúmenes de datos, posibilitando a las organizaciones tomar decisiones más informadas (López García, 2017).

Uno de los sectores que encara desafíos significativos en la gestión y mantenimiento es el sector vial, especialmente en áreas rurales. En la Provincia de Santo Domingo de los Tsáchilas, por ejemplo, la red vial se extiende por más de 2.953,86 km, con una composición diversa: 11% de asfalto, 8% de tierra y un predominante 83% de lastre (SIL G. P., 2023).

Esta diversidad en las carreteras plantea un conjunto único de desafíos para el Gobierno Provincial, cuya capacidad operativa actual para el mantenimiento vial es insuficiente, especialmente durante la época invernal. Esta limitación se refleja en las numerosas quejas por parte de los ciudadanos y resalta la necesidad de soluciones innovadoras.

Adicionalmente, la provincia se encuentra en una zona de alta vulnerabilidad a desastres naturales como movimientos de masa, inundaciones y deslizamientos, especialmente durante las fuertes épocas invernales. Estas condiciones de riesgo amplifican la importancia de una planificación y ejecución efectiva del mantenimiento vial, que debe ser ágil para responder a situaciones emergentes.

En este marco, la implementación de un algoritmo de ML de aprendizaje no supervisado para la creación de un cronograma de mantenimiento vial en las vías rurales de la Provincia de Santo Domingo de los Tsáchilas no solo tiene el potencial de mejorar significativamente la eficiencia y eficacia del mantenimiento vial, sino también podría mejorar la seguridad y calidad de vida de los habitantes de la provincia.

Según el último Censo de Población y Vivienda 2022, Ecuador cuenta con 16.938.986 habitantes y, en la provincia en cuestión reside una población total de 492.969 habitantes, con un 24.9% que habita en la zona rural, distribuidos en sus 2 cantones, y sus 8 parroquias urbanas y 10 rurales (INEC, 2023).

La aplicación de ML en el mantenimiento preventivo está revolucionando la industria, anticipándose a las fallas de los equipos y reduciendo costos para las empresas. En el ámbito del mantenimiento industrial, el ML permite la detección de anomalías incipientes basada en aprendizaje automático, con el objetivo de predecir fallas en la maquinaria, de manera que las reparaciones puedan ser programadas sin interrumpir el proceso de producción (Nina, 2022). Esta innovación es crucial en el contexto de las vías rurales de Santo Domingo de los Tsáchilas, donde el mantenimiento proactivo puede

contribuir a la reducción de los riesgos asociados con los desastres naturales y mejorar la calidad de las infraestructuras viales.

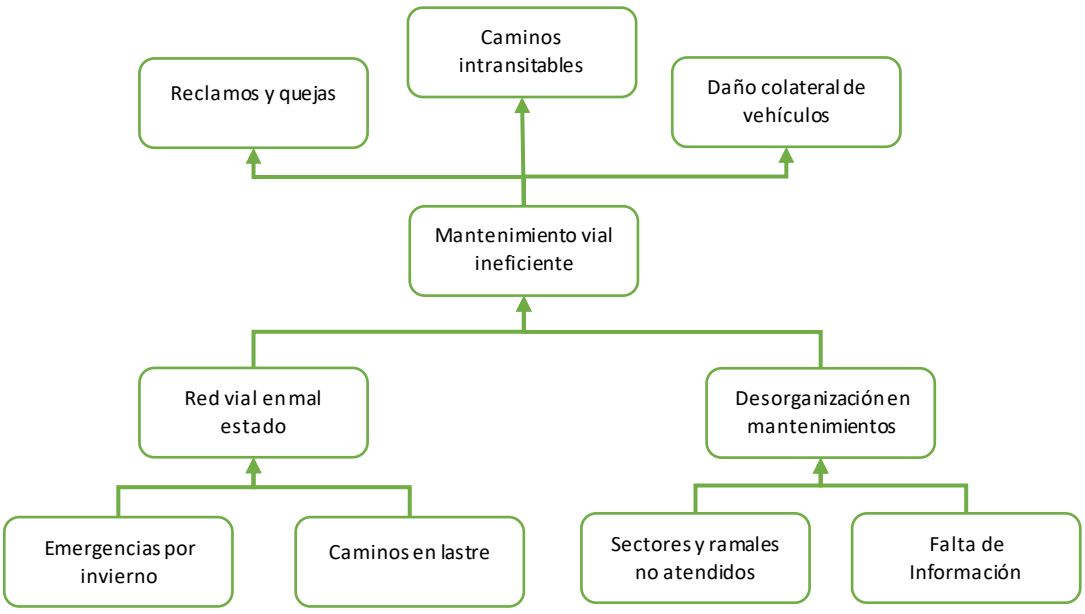
1.3 Planteamiento del Problema

La Provincia de Santo Domingo de los Tsáchilas, pese a su relativamente reciente existencia jurisdiccional de 16 años, cuenta con más de 132 años de historia territorial y una infraestructura vial diversa y compleja. Esta diversidad se manifiesta robustamente en su red vial rural, compuesta por caminos de asfalto, tierra y lastre, predominando este último material.

La comunidad local ha expresado crecientemente su descontento con el estado de la infraestructura vial, especialmente en lo que respecta a la capacidad del Gobierno Provincial para mantenerla en condiciones óptimas, sobre todo durante las épocas invernales.

En estas temporadas, el rápido deterioro de los caminos rurales desafía la capacidad instalada del Gobierno para efectuar mantenimientos rutinarios y atender situaciones emergentes. Según el marco legal vigente, la responsabilidad de la intervención y mantenimiento de la red vial rural recae en el Gobierno Provincial para los cantones de Santo Domingo y La Concordia.

Figura 1. Árbol de Problemas, Estado Sistema Vial



Fuente: Pérez M, 2023

Este desafío se magnifica por la ausencia de un sistema de monitoreo eficiente que posibilite una intervención proactiva antes y durante la etapa invernal. La falta de tal mecanismo no solo repercute negativamente en la calidad del servicio proporcionado a la comunidad, sino que también plantea riesgos de seguridad y bienestar para los residentes de la provincia.

Ante la complejidad y la magnitud del problema, se evidencia la necesidad de explorar soluciones innovadoras que trasciendan las estrategias convencionales de mantenimiento. En este escenario, la implementación de un algoritmo de ML de aprendizaje no supervisado emerge como una oportunidad inigualable para optimizar la planificación y ejecución del mantenimiento vial, prometiendo una intervención más eficaz y eficiente que responda tanto a las exigencias de la comunidad como a las condiciones ambientales adversas que enfrenta la provincia.

1.4 Objetivos de la Investigación

1.4.1 Objetivo General:

Implementar un algoritmo de *Machine Learning* de aprendizaje no supervisado para optimizar la planificación del mantenimiento vial en la red rural de la provincia de Santo Domingo de los Tsáchilas.

1.4.2 Objetivos Específicos:

- Diseñar e implementar un algoritmo de aprendizaje no supervisado que utilice datos históricos y actuales para identificar patrones y tendencias en el estado de la red vial rural.
- Evaluar la eficacia del algoritmo implementado mediante la utilización de métricas de rendimiento y ajustar el modelo según los resultados.
- Generar recomendaciones específicas para la gestión y mantenimiento de la red vial rural, basadas en los resultados del algoritmo, e identificar oportunidades para una planificación del uso del suelo más sostenible.

1.5 Alcance

El presente trabajo de titulación se enfoca en la implementación de técnicas de aprendizaje no supervisado para optimizar la gestión del mantenimiento de las vías rurales en Santo Domingo de los Tsáchilas. Utilizando métodos de aprendizaje automático.

El proyecto tiene como objetivo identificar patrones y características en el estado y uso de las vías rurales. Este análisis permitirá una comprensión más profunda de los desafíos y necesidades específicas de la red vial de la provincia.

La metodología CRISP-DM orientará el desarrollo del proyecto, asegurando un enfoque estructurado desde la comprensión inicial del problema hasta la implementación de soluciones analíticas y predictivas. Esta metodología estándar en la minería de datos brindará una base sólida para un análisis detallado y la implementación de sistemas analíticos automatizados.

El proyecto incluirá una revisión de la literatura científica y técnica relacionada con métodos no supervisados aplicables en contextos similares, lo que ayudará a seleccionar los algoritmos más adecuados y mejorar los resultados.

Dado que el proyecto se centra en las vías rurales de Santo Domingo de los Tsáchilas, los resultados y recomendaciones estarán condicionados a los datos y recursos disponibles. Se espera que este trabajo aporte *insights* valiosos y estrategias de mejora en la gestión vial, contribuyendo a la seguridad y eficiencia en el mantenimiento de las carreteras de la provincia.

Capítulo 2

Marco Teórico

Este capítulo establece el marco teórico y contextual de la investigación. Se abordarán conceptos esenciales en aprendizaje automático y mantenimiento vial, y se describirá la red vial rural en los cantones de Santo Domingo y La Concordia. Se incluirá la tipología de vías y su estado actual para identificar desafíos y oportunidades en la gestión del mantenimiento vial.

2.6 Ciencia de Datos

La ciencia de datos se centra en la inferencia estadística, la predicción y la toma de decisiones a través de métodos analíticos avanzados. No se limita a la recopilación y análisis de datos, sino que también incluye la construcción y validación de modelos predictivos y prescriptivos, así como la visualización y la comunicación de resultados (Cleveland, 2001) para informar decisiones. Sin embargo, no abarca la recolección de datos sin un análisis posterior ni la toma de decisiones no basada en evidencia (Cleveland, 2001).

2.7 Inteligencia Artificial

La Inteligencia Artificial (IA) se define como la simulación de procesos de inteligencia humana por sistemas informáticos. Estos procesos incluyen el aprendizaje (la adquisición de información y las reglas para usar la información), el razonamiento (usar las reglas para alcanzar conclusiones aproximadas o definitivas) y la autocorrección. En (Russell, 2016) se ofrece una perspectiva amplia al clasificar la IA en cuatro categorías: sistemas que piensan como humanos, sistemas que actúan como humanos, sistemas que piensan racionalmente y sistemas que actúan racionalmente.

2.8 Machine Learning

El *Machine Learning*, es una rama de la inteligencia artificial que ha evolucionado desde la creación de programas simples que realizaban tareas específicas, hasta sistemas complejos capaces de aprender y adaptarse. El término fue acuñado por Arthur Samuel

en 1959, refiriéndose al campo de estudio que da a las computadoras la capacidad de aprender sin estar explícitamente programadas. (Mitchell, 1997) ofrece una definición formal: "Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su rendimiento en las tareas en T, medido por P, mejora con la experiencia E".

2.8.1 Principales Paradigmas del Machine Learning

El ML se categoriza comúnmente en tres paradigmas principales: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. El aprendizaje supervisado utiliza datos etiquetados para predecir resultados, mientras que el aprendizaje no supervisado busca patrones en datos no etiquetados. El aprendizaje por refuerzo se centra en cómo los agentes inteligentes deben tomar acciones en un entorno para maximizar la noción de recompensa acumulativa. (Alpaydin, 2020) proporciona una visión detallada de estos paradigmas y sus aplicaciones en el mundo real.

2.8.2 Aplicaciones Actuales y Potencial del Machine Learning

Las aplicaciones del ML son vastas y afectan prácticamente a todos los sectores de la industria. En la salud, se utiliza para predecir enfermedades y personalizar tratamientos. En el comercio, impulsa sistemas de recomendación personalizados. En la gestión de infraestructuras, el ML puede predecir el desgaste de los materiales y programar mantenimientos preventivos. Jordan y Mitchell (Jordan, 2015) discuten cómo el ML está siendo integrado en sistemas complejos para mejorar la toma de decisiones y automatizar procesos.

2.8.3 Tipos de Aprendizaje en *Machine Learning*

2.8.3.1 Aprendizaje Supervisado

El aprendizaje supervisado es uno de los enfoques más comunes en el campo del *Machine Learning*. En este paradigma, se alimenta al algoritmo con datos de entrada etiquetados, y el modelo aprende a predecir las etiquetas a partir de nuevas instancias no vistas. Los algoritmos típicos incluyen la regresión lineal y logística, máquinas de soporte vectorial (SVM), árboles de decisión y redes neuronales. En (Hastie, 2009) se describe cómo estos algoritmos se utilizan para clasificar datos o predecir valores continuos.

Las aplicaciones del aprendizaje supervisado son amplias y abarcan desde la clasificación de correos electrónicos en spam o no spam, hasta la predicción de la demanda de productos en el inventario. Un ejemplo relevante podría ser la utilización de estos algoritmos para predecir el tiempo restante útil (RUL) de los activos en mantenimiento vial.

2.8.3.2 Evaluación y Validación de Modelos

La evaluación y validación de modelos son pasos cruciales en el aprendizaje supervisado para asegurar que los modelos funcionen bien con datos nuevos y no vistos. Esto implica utilizar técnicas como la validación cruzada y métricas de rendimiento como la precisión, la curva ROC y el área bajo la curva (AUC), el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). En (James G. W., 2013) se destaca la importancia de la validación de modelos para evitar problemas como el sobreajuste, garantizando que el modelo generalice bien a partir de los datos de entrenamiento.

2.8.3.3 Aprendizaje No Supervisado

El aprendizaje no supervisado involucra la modelación de patrones y estructuras en datos que no están etiquetados. Este paradigma de aprendizaje automático es esencial cuando no se conocen las respuestas correctas de antemano. Los algoritmos populares dentro de este enfoque incluyen el *clustering*, como K-means y DBSCAN, y la reducción de dimensionalidad, como el análisis de componentes principales (PCA) y el t-SNE. (Murphy, 2012) proporciona una visión comprensiva de estas técnicas y sus fundamentos matemáticos.

Las técnicas de aprendizaje no supervisado son fundamentales para entender la estructura intrínseca de los datos y descubrir patrones subyacentes sin la influencia de una variable objetivo predefinida. Estas técnicas son especialmente valiosas para explorar datos, realizar segmentación de mercado y detectar anomalías.

2.8.3.4 Casos de Uso y Beneficios

El aprendizaje no supervisado tiene una amplia gama de aplicaciones prácticas en diversas industrias. En el sector comercial, se utiliza para la segmentación de clientes y el análisis de cestas de mercado. En biología, ayuda en la genómica para agrupar genes

con funciones similares. En el mantenimiento vial, puede identificar patrones de deterioro o clasificar tipos de carreteras basándose en características no etiquetadas.

Uno de los principales beneficios del aprendizaje no supervisado es su capacidad para manejar datos sin etiquetar, lo que es común en muchos contextos del mundo real. Además, puede revelar relaciones inesperadas en los datos, lo que lleva a nuevos *insights* y descubrimientos. (Xie, 2016) señala cómo las técnicas no supervisadas pueden proporcionar una visión significativa cuando las etiquetas de datos son escasas o inexistentes.

2.8.4 Algoritmos de ML

Los algoritmos de *Machine Learning* (ML) se clasifican según el tipo de aprendizaje y la función que desempeñan. La clasificación incluye aprendizaje supervisado, no supervisado, semi-supervisado y por refuerzo. Dentro de estas categorías, los algoritmos se dividen más específicamente en clasificación, regresión, *clustering* y reducción de dimensionalidad, entre otros. (Bishop, 2006) proporciona una taxonomía detallada de los algoritmos de ML y las teorías subyacentes que guían su desarrollo y aplicación.

2.8.4.1 Algoritmos de Clustering

Los algoritmos de *clustering* agrupan conjuntos de datos basándose en la similitud entre los miembros de un conjunto. Los más conocidos son, por un lado, *K-means*, que divide los datos en K grupos basándose en la proximidad a los centroides de los clusters; por otro lado, DBSCAN, que identifica regiones de alta densidad y las separa de regiones de baja densidad.

Estos métodos son fundamentales en la detección de patrones naturales en los datos y tienen aplicaciones que van desde la segmentación de clientes hasta la organización de grandes bibliotecas de documentos. (Jain, 2010) explora las fortalezas y limitaciones de diversos métodos de *clustering* en diferentes contextos.

2.8.4.2 Reducción de Dimensionalidad

La reducción de dimensionalidad busca simplificar los datos sin perder información importante. Técnicas como el Análisis de Componentes Principales (PCA) y el Análisis de Componentes Independientes (ICA) son utilizadas para reducir el número de variables en el análisis y mejorar la eficiencia de otros algoritmos de ML. Estas técnicas son

particularmente útiles cuando se trata con datos de alta dimensionalidad y pueden ayudar a mejorar tanto la interpretación como la visualización de los datos. Roweis y Saul (Roweis, 2000) describen cómo la reducción de dimensionalidad puede ser utilizada para descubrir la estructura subyacente de los datos.

2.8.4.3 Algoritmos de Asociación

Los algoritmos de asociación, como el Apriori y el Eclat, son utilizados para encontrar relaciones entre variables en grandes bases de datos. Estos son comúnmente usados en el análisis de cestas de mercado para identificar reglas de asociación que pueden prever el comportamiento del consumidor. Estos algoritmos son fundamentales en el comercio minorista para el diseño de estrategias de *marketing* y colocación de productos. Agrawal y Srikant (Agrawal, 1994) demuestran cómo las reglas de asociación pueden ser eficazmente utilizadas para descubrir patrones en transacciones de datos a gran escala.

2.8.5 Minería de Datos y Extracción de Conocimiento

La minería de datos es el proceso de descubrir patrones y conocimientos a partir de grandes volúmenes de datos. Se basa en métodos de ML, estadística y sistemas de bases de datos para extraer información útil o interesante que no se podría obtener mediante una simple inspección. En (Fayyad, 1996) se establecieron los principios fundamentales de la minería de datos, delineando el proceso de conocimiento en bases de datos (KDD) que incluye la selección, el preprocesamiento, la transformación, la minería de datos y la interpretación/evaluación.

2.8.5.1 Análisis Exploratorio de Datos

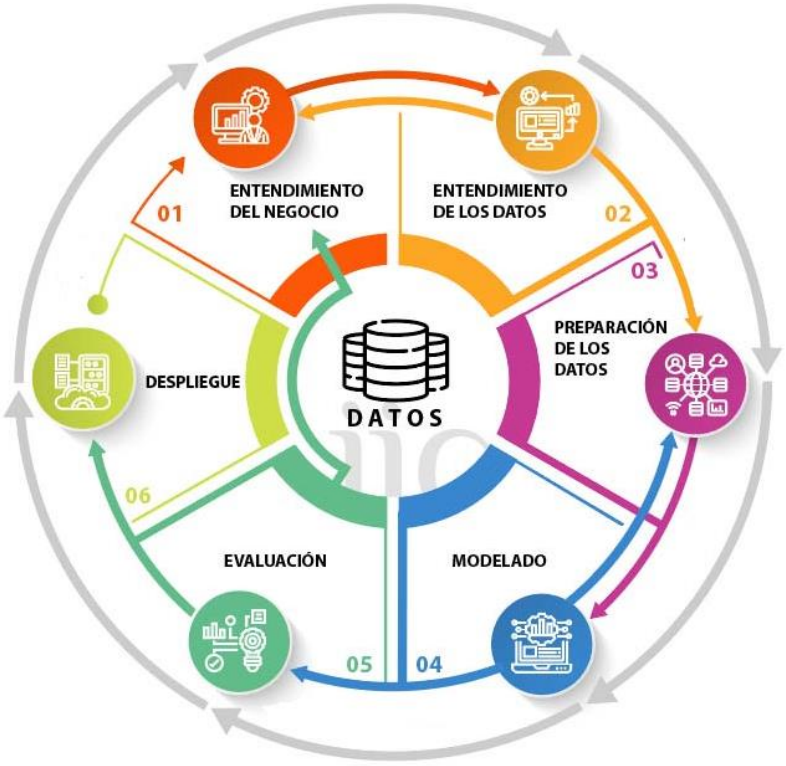
El Análisis Exploratorio de Datos (EDA) es una aproximación crítica en la minería de datos que utiliza técnicas gráficas y cuantitativas para maximizar el *insight* en un conjunto de datos, descubrir estructuras subyacentes, extraer variables importantes, detectar *outliers* y anomalías, y probar hipótesis subyacentes. (Tukey, 1977) introdujo el EDA como una filosofía de análisis de datos que promueve la exploración y la apertura a nuevas posibilidades, sin la restricción de hipótesis preconcebidas.

2.8.5.2 Preprocesamiento de Datos

El preprocesamiento de datos es un paso crucial en la minería de datos que involucra la preparación y transformación de datos crudos en un formato adecuado para el análisis posterior. Incluye la limpieza de datos, la integración, la transformación, la reducción y la discretización. La limpieza se ocupa de manejar datos faltantes y ruidosos, la integración implica la combinación de datos de múltiples fuentes, y la transformación y reducción buscan derivar atributos más útiles y reducir la dimensionalidad de los datos. En (García, 2015) se discuten diversas técnicas de preprocesamiento y su impacto en la calidad de los resultados de minería de datos.

2.8.6 Metodología CRISP-DM para el Desarrollo de Proyectos de ML

Figura 2. Ciclo de la Metodología CRISP-DM.



Fuente: Pérez M, 2023

2.8.6.1 Entendimiento del Negocio

Esta fase inicial del proceso CRISP-DM implica comprender el proyecto desde una perspectiva de negocio y definir los objetivos del proyecto de ML en términos de negocio. Se debe realizar una evaluación de la situación actual, identificar los requerimientos de

negocio, los criterios de éxito y los riesgos potenciales. Wirth y Hipp (Wirth, 2000) describen cómo el entendimiento del negocio guía la dirección del proyecto de ML y asegura que los resultados sean relevantes y valiosos para la organización.

2.8.6.2 Entendimiento de los Datos

El entendimiento de los datos implica recoger los datos iniciales, familiarizarse con ellos, identificar problemas de calidad de datos, y descubrir los primeros *insights*. Se realiza una evaluación preliminar para entender la estructura de los datos, las propiedades de las variables importantes y la distribución de las variables clave. En (Chapman, 2000) se explican cómo esta comprensión inicial es esencial para preparar adecuadamente los datos y seleccionar las técnicas analíticas adecuadas.

2.8.6.3 Preparación de los Datos

La preparación de los datos es una de las fases más críticas y que consume más tiempo en CRISP-DM. Incluye la limpieza de datos, la creación de variables derivadas, la integración de datos y la transformación de formatos. El objetivo es desarrollar un conjunto de datos final para modelar que sea de alta calidad y esté estructurado adecuadamente para el análisis. (Pyle, Data Preparation for Data Mining, 1999) destaca la importancia de una preparación de datos meticulosa para el éxito de los proyectos de minería de datos y ML.

2.8.6.4 Modelado

En la fase de modelado, se seleccionan y aplican diversas técnicas de modelado con los datos preparados. Se deben considerar varios modelos y técnicas, ajustando los parámetros del modelo para optimizar la calidad de las predicciones o clasificaciones. (Shearer, 2000) señala que la selección del modelo correcto depende tanto de la naturaleza del problema como de la calidad de los datos.

2.8.6.5 Evaluación

La fase de evaluación consiste en determinar si el modelo cumple con los objetivos del negocio y si hay algún problema desde una perspectiva de negocio. Se evalúa el desempeño del modelo utilizando un conjunto de datos de prueba y se revisa el proceso para asegurar que se han considerado todos los factores importantes. Se deben tomar decisiones sobre la posible implementación del modelo en el entorno operativo.

2.8.6.6 Despliegue

El despliegue es la finalización del proceso CRISP-DM, donde el modelo de ML se implementa en el entorno operativo para tomar decisiones en tiempo real o se entregan los hallazgos y conocimientos a los interesados. El despliegue puede variar desde la generación de un informe hasta la implementación de un proceso automatizado y recurrente.

2.9 Conceptos Viales y Mantenimiento

2.9.1 Mantenimiento Vial

El mantenimiento vial es una parte esencial de la gestión de infraestructuras de transporte, crucial para la seguridad, la eficiencia y la economía. La inversión en mantenimiento regular de las vías no solo previene el deterioro acelerado y reduce la necesidad de reparaciones costosas a largo plazo, sino que también asegura la seguridad de los usuarios y la fluidez del tráfico, lo que es esencial para el desarrollo económico. Según la Asociación Mundial de la Carretera (PIARC), un mantenimiento vial adecuado contribuye directamente a la reducción de accidentes y al ahorro en costos operativos vehiculares (PIARC, 2016).

2.9.1.1 Tipos de Mantenimiento: Preventivo, Correctivo y Predictivo

Preventivo

El mantenimiento preventivo se realiza regularmente, independientemente del estado actual de la vía, para prevenir la aparición de problemas. Incluye actividades como la limpieza de cunetas, la poda de vegetación y el sellado de grietas, con el objetivo de prolongar la vida útil de la carretera y reducir la probabilidad de fallos.

Correctivo

El mantenimiento correctivo se refiere a las actividades realizadas en respuesta a problemas ya identificados, como baches o daños en la superficie del pavimento. Este tipo de mantenimiento es a menudo más costoso y puede llevar a interrupciones del tráfico mientras se realizan las reparaciones.

Predictivo

El mantenimiento predictivo, potenciado por el avance de la ciencia de datos y el ML, consiste en utilizar datos históricos y en tiempo real para prever cuándo se necesitarán intervenciones de mantenimiento antes de que los problemas se vuelvan evidentes. El objetivo es programar las actividades de mantenimiento de forma eficiente, minimizando los costos y las interrupciones del tráfico. Shalaby y Farhan (Shalaby, 2004) discuten cómo la implementación de sistemas de gestión de mantenimiento predictivo puede optimizar los recursos y mejorar la planificación de las actividades de mantenimiento.

2.10 Composición y Estructura de las Vías Rurales

Las vías rurales constituyen una infraestructura crítica que soporta las actividades económicas y sociales en áreas fuera de los centros urbanos. Su composición y estructura están influenciadas por factores geográficos, económicos y materiales disponibles (Haas, 2003).

2.10.1 Tipos de Vías y Materiales Utilizados

Las vías rurales generalmente se clasifican según los materiales utilizados en su construcción y pueden ser de asfalto, tierra o lastre (Pieplow, 1995).

2.10.1.1 Asfalto

El asfalto es ampliamente utilizado por su durabilidad y capacidad para proporcionar una superficie lisa, lo que resulta en menor desgaste de los vehículos y mayor comodidad para los conductores. Las innovaciones en materiales asfálticos, como la incorporación de caucho triturado, han mejorado las propiedades de estos pavimentos (Smith, 2001).

2.10.1.2 Tierra

Las carreteras de tierra son comunes en regiones donde la inversión en infraestructura es limitada, pero requieren un mantenimiento constante para asegurar la transitabilidad y seguridad (Pieplow, 1995).

2.10.1.3 Lastre

Las carreteras de lastre, compuestas por capas de grava compactada, son una solución intermedia entre el asfalto y las carreteras de tierra, proporcionando una opción de menor costo y mantenimiento que el asfalto, pero con mejor estabilidad que las carreteras de tierra pura (Pieplow, 1995).

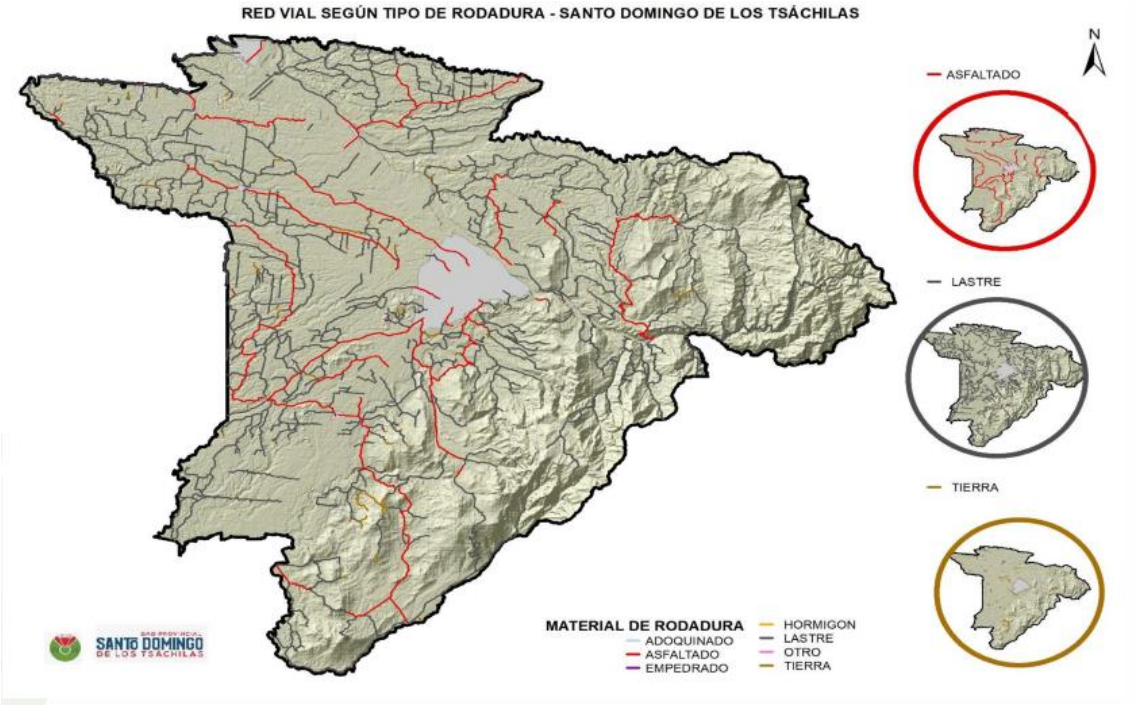
2.10.2 Capas de Rodadura y su Comportamiento

La capa de rodadura es la superficie sobre la cual viajan los vehículos y es fundamental para la seguridad. Su diseño debe considerar la textura y resistencia al deslizamiento para prevenir accidentes, especialmente en condiciones meteorológicas adversas (Thompson, 1979).

2.10.3 Red vial provincial.

Santo Domingo de los Tsáchilas posee características de movilidad con una vocación dirigida al comercio, que se manifiesta en el tránsito vehicular, en sus vías y terminales de transporte como ejes conectores. Estos equipamientos de transporte dedicados al transporte interprovincial e inter parroquial tienen un impacto significativo dentro de la dinámica de la ciudad, especialmente en las vías urbanas, en la conectividad y movilidad interna de las personas, contamos con un sistema vial provincial rural de un total de 2.953,86 Km presentadas en el (SIL S. d., 2020)

Figura 3. Capas de Rodadura Provincia de Santo Domingo de los Tsáchilas



Fuente: GADPSDT, 2020

2.10.4 Mantenimiento vial provincial

El GAD Provincial ofrece dos programas de mantenimiento anual: el de lastre y asfalto. Un equipo técnico y operativo brinda este mantenimiento de manera anual mediante una programación de intervención por parroquia. En algunos casos, debido a las condiciones climáticas, este mantenimiento se realiza dos veces al año.

El Gobierno Autónomo Descentralizado (GAD) Provincial realiza mantenimiento anual en aproximadamente 1.000 km de vías de segundo orden, especialmente las de capa de rodadura de lastre y tierra que conectan varios centros poblados rurales. La red vial provincial consta de 2953.86 km, de los cuales el 82.6% es lastrado, 6.1% es de tierra y 11.0% está pavimentado. Según el Plan de Desarrollo y Ordenamiento Territorial (PDyOT, 2020), esta cifra indica que anualmente no todas las vías rurales de la provincia reciben mantenimiento frecuente, lo que implica un déficit de mantenimiento del 46.7%. Para cubrir la demanda de mantenimiento en toda la provincia de manera más frecuente, se debería triplicar la cantidad de maquinaria, como retroexcavadoras, rodillos, tanqueros y motoniveladoras.

2.10.4.1 Maquinaria de mantenimiento vial.

La Provincia Tsáchila cuenta con una flota de 72 maquinarias que permiten realizar mantenimiento y mejoramiento de las vías de la provincia. La maquinaria de mantenimiento vial está conformada de camiones, cargadores, carro taller, excavadoras, minicargadores, motoniveladoras, plataformas, rodillos, taqueros, tractores, triturador y en su mayoría volquetes. Referente al estado de la maquinaria se ha encontrado que el 39% no están en condiciones de operatividad según el informe pasado por el departamento de Obras Públicas (OOPP) del GAD (PDyOT, 2020) .

2.10.4.2 Cronograma de mantenimiento vial.

Es un plan detallado que se utiliza para llevar a cabo mantenimientos en una red de carreteras o vías de transporte en un período de tiempo determinado. Este tipo de cronograma permite planificar y programar las tareas de mantenimiento de manera oportuna y eficiente, lo que contribuye a mantener el buen estado de las carreteras y garantizar la seguridad de los usuarios.

Además, el cronograma de mantenimiento vial también se utiliza para llevar un registro de los mantenimientos realizados y para hacer un seguimiento del cumplimiento de los plazos establecidos. Esto permite a las autoridades responsables del mantenimiento vial monitorear el desempeño y determinar si se están cumpliendo los objetivos establecidos en cuanto a la calidad y la eficiencia del mantenimiento.

2.10.5 Análisis del Sistema Vial Rural.

En el inventario de vías de Santo Domingo de los Tsáchilas de 2017, se detalla que el 55,94% de las vías se ubican en tierras agrícolas, un 37,50% en zonas ganaderas, un 3,80% en suelos improductivos y el 2,40% restante conduce a áreas pobladas. De los 2953.86 km que comprende la red vial provincial rural, apenas el 2.71% se encuentra en buen estado.

La red vial estatal abarca 208,69 km, mientras que 50,26 km corresponden a vías urbanas que se conectan con la red rural. El mantenimiento de estas vías se programa anualmente por parroquia, y en algunos casos, debido a las condiciones climáticas de la región, se requiere más de una intervención al año. La infraestructura incluye 2.057

alcantarillas y 387 puentes, además de 71 puentes badén. Para mejorar la conectividad en las zonas rurales, se reconoce la necesidad de construir más puentes.

Capítulo 3

Este capítulo describe la implementación de un algoritmo de aprendizaje no supervisado para mejorar el cronograma de mantenimiento vial. A través de la metodología CRISP-DM, se llevará a cabo la recolección y análisis de datos históricos y actuales, con el objetivo de identificar patrones clave en el estado de las vías rurales. La aplicación de este enfoque posibilitará la creación de un modelo avanzado que optimice la gestión y planificación del mantenimiento vial.

Metodología.

Para este proyecto, se ha elegido la metodología CRISP-DM, un estándar reconocido en el ámbito de la minería de datos, debido a su estructura bien definida y eficacia probada. La metodología CRISP-DM será aplicada en el análisis y optimización de la gestión del mantenimiento de las vías rurales de Santo Domingo de los Tsáchilas. Este marco metodológico se adaptará para aplicar técnicas de aprendizaje no supervisado, con el objetivo de identificar patrones y tendencias significativas en el estado y uso de las vías rurales, así como en su mantenimiento.

3.1 Entendimiento del negocio

3.1.1 Definición de Objetivos del Proyecto

El objetivo primordial de este proyecto es la creación de un cronograma de mantenimiento vial altamente confiable para la red de carreteras rurales en la provincia de Santo Domingo de los Tsáchilas, enfocándonos en las vías de lastre. Este cronograma se basará en el análisis de datos históricos y actuales, aplicando técnicas avanzadas de aprendizaje no supervisado. La meta es identificar con precisión las áreas más críticas que requieren intervención, priorizando las acciones de mantenimiento de manera efectiva.

3.1.2 Relevancia del Mantenimiento Vial y Necesidades Actuales

La red vial de Santo Domingo y La Concordia, caracterizada por su variada tipología que incluye asfalto, tierra y lastre, enfrenta desafíos significativos, especialmente durante las temporadas de lluvia. El rápido deterioro de las vías rurales durante estas épocas pone

a prueba la capacidad del Gobierno Provincial para mantener y restaurar estas carreteras de manera oportuna. Un enfoque proactivo, utilizando la ciencia de datos para monitorear y predecir el deterioro vial, se presenta como una solución vital. Este enfoque no solo mejorará la eficiencia del mantenimiento, sino que también reducirá las quejas de los ciudadanos y mejorará la seguridad y accesibilidad en la provincia.

3.1.3 Objetivos y Criterios de Éxito del Negocio

El éxito del proyecto se medirá por la eficacia del cronograma de mantenimiento desarrollado. Un cronograma exitoso deberá ofrecer recomendaciones precisas sobre qué vías deben ser atendidas con prioridad, asegurando un alto grado de fiabilidad y satisfacción ciudadana. Además, un criterio clave de éxito será la capacidad del cronograma para reducir las quejas de los ciudadanos mediante una respuesta más rápida y eficaz a los problemas viales.

3.1.4 Evaluación de la Situación

Se dispone de una base de datos detallada que abarca las características y condiciones de las vías rurales en la provincia de Santo Domingo de los Tsáchilas. Esta base de datos incluye una amplia gama de información relevante, recopilada a lo largo de varios años, lo que asegura que contamos con un volumen de datos suficiente para abordar nuestro problema.

La información recogida abarca aspectos como la longitud y el ancho de las vías, el tipo de terreno, el clima predominante en la zona, el estado actual de la vía, y otros datos relevantes como el tráfico promedio, la presencia de curvas y puentes, y la producción económica asociada a las zonas servidas por estas vías. Esta información no solo es cuantitativa, sino también cualitativa, proporcionando un panorama completo de las condiciones y desafíos asociados a cada segmento de la red vial.

3.2 Comprensión de los Datos

En esta fase se examina la base de datos de la red vial para evaluar su calidad y entender su estructura, lo que es fundamental para identificar patrones clave y formular hipótesis iniciales, siguiendo la metodología CRISP-DM.

3.2.1 Recolectar los Datos Iniciales para el Proyecto de Mantenimiento Vial

Para este proyecto, se han utilizado datos específicos relativos a la red vial rural de la provincia de Santo Domingo de los Tsáchilas. Estos datos incluyen información detallada sobre las características y condiciones de las vías, lo que es esencial para realizar un análisis efectivo y generar predicciones precisas. A diferencia del ejemplo, donde se utilizaban datos ficticios, en nuestro caso, los datos son reales y han sido obtenidos directamente de fuentes gubernamentales o autoridades locales de infraestructura vial.

Los datos adquiridos incluyen:

Características de la Vía: Información detallada sobre cada segmento de la red vial, como longitud, anchura, tipo de superficie, etc.

Condiciones Ambientales y de Uso: Datos sobre el clima, el tipo de terreno, el tráfico, y otros factores relevantes.

Historial de Mantenimiento: Registros de mantenimientos anteriores y el estado actual de las vías.

Los atributos específicos que serán cruciales para el análisis son:

- Identificadores geográficos de las vías (provincia, cantón, parroquia).
- Medidas físicas y técnicas de las vías (longitud, ancho, tipo de terreno).
- Datos sobre el uso y el estado de las vías (tráfico, número de curvas, estado vial).

La base de datos está organizada en una sola tabla que refleja diferentes aspectos de la red vial, permitiendo un análisis integral y multifacético de las condiciones y necesidades de mantenimiento.

Figura 4. Tipos de datos presentes en la tabla

```
# Utilizaremos el método .info() para obtener información
selected_data_info = selected_data.info()
selected_data_info
```

```
STDOUT/STDERR
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 756 entries, 0 to 755
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   codunico        756 non-null    float64
 1   canton          756 non-null    object
 2   parr            756 non-null    object
 3   clima           756 non-null    object
 4   terreno_tipo    756 non-null    object
 5   carrilles       756 non-null    object
 6   estado_vial     756 non-null    object
 7   rodea_via       756 non-null    object
 8   ancho_mt        756 non-null    float64
 9   longitud_km     756 non-null    float64
 10  meses_manten    756 non-null    int64
 11  estado_via      756 non-null    int64
 12  velocidad_prom  756 non-null    float64
 13  ncurvas         756 non-null    int64
 14  npuentes        756 non-null    int64
 15  tpda            756 non-null    int64
 16  produccion_usd  756 non-null    float64
 17  nviviendas      756 non-null    int64
 18  poblacion       756 non-null    int64
dtypes: float64(5), int64(7), object(7)
memory usage: 112.3+ KB
```

3.2.2 Descripción de los Datos

Se examina y detalla la estructura y contenido de la base de datos proporcionada. Esta base de datos almacena información crucial sobre la red vial de la provincia de Santo Domingo de los Tsáchilas y está diseñada para soportar el análisis enfocado en optimizar el cronograma de mantenimiento de las vías.

La tabla única con la que contamos compila un conjunto de 19 variables que caracterizan cada segmento vial comprendidos en 756 registros. Entre los atributos más relevantes se encuentran la identificación geográfica, las dimensiones físicas de las vías, el tipo de terreno, las condiciones climáticas, y el estado actual de la infraestructura. Estos datos se registran en formatos numéricos y alfanuméricos, adecuados para aplicar métodos de aprendizaje no supervisado y otras técnicas de análisis estadístico.

Información de atributos:

Tabla 3. Atributos tabla de datos Red Vial Rural Lastre Santo Domingo de los Tsáchilas

Orden	*Variable*	*Descripción*
1	codunico	Código único asignado a cada segmento de la vía
2	canton	Cantón al que pertenece el segmento de la vía
3	parr	Parroquia en la que se encuentra el segmento de la vía
4	clima	Clima predominante en la zona del segmento de la vía
5	terreno_tipo	Tipo de terreno sobre el cual está construida la vía
6	carrilles	Número de carriles que tiene el segmento de la vía
7	estado_vial	Estado actual del segmento de la vía
8	rodea_via	Entorno que rodea al segmento de la vía
9	ancho_mt	Ancho del segmento de la vía medido en metros
10	longitud_km	Longitud del segmento de la vía medida en kilómetros
11	meses_manten	Frecuencia de mantenimiento del segmento de la vía en meses
12	estado_via	Código que indica el estado de la vía
13	velocidad_prom	Velocidad promedio estimada para el tránsito en el segmento de la vía
14	ncurvas	Número de curvas presentes en el segmento de la vía
15	npuentes	Número de puentes presentes en el segmento de la vía
16	tpda	Tráfico promedio diario anual en el segmento de la vía
17	produccion_usd	Valor de la producción económica en USD asociada al área de influencia de la vía
18	nviviendas	Número de viviendas en el área de influencia del segmento de la vía
19	poblacion	Población estimada en el área de influencia del segmento de la vía

3.2.3 Exploración de los Datos

En la fase de exploración de datos, se ha llevado a cabo un análisis estadístico detallado para comprender mejor las características y condiciones de la red vial. Este análisis abarca 756 segmentos viales y proporciona medidas de tendencia central y dispersión para diversas variables.

Tabla 4. Medidas de Tendencia central y Dispersión

Variables	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
ancho_mt	3.64	1.04	1.95	2.88	3.41	4.15	7.91
longitud_km	2.49	2.57	0.00	0.93	1.82	3.17	19.69
meses_manten	13.04	2.96	1.00	11.00	13.00	15.00	22.00
velocidad_prom	23.32	7.03	9.89	18.16	22.02	28.02	43.48
ncurvas	9.88	11.29	0.00	3.00	7.00	13.00	84.00
npuentes	0.38	0.69	0.00	0.00	0.00	1.00	4.00
tpda	4.76	7.21	0.00	0.00	0.00	8.00	39.00
produccion_usd	17,090.95	20,239.11	0.00	4,579.98	11,748.70	22,959.40	172,059.20
nviviendas	38.27	27.30	6.00	17.00	29.00	55.00	155.00
poblacion	151.49	108.29	24.00	67.00	117.00	216.50	603.00

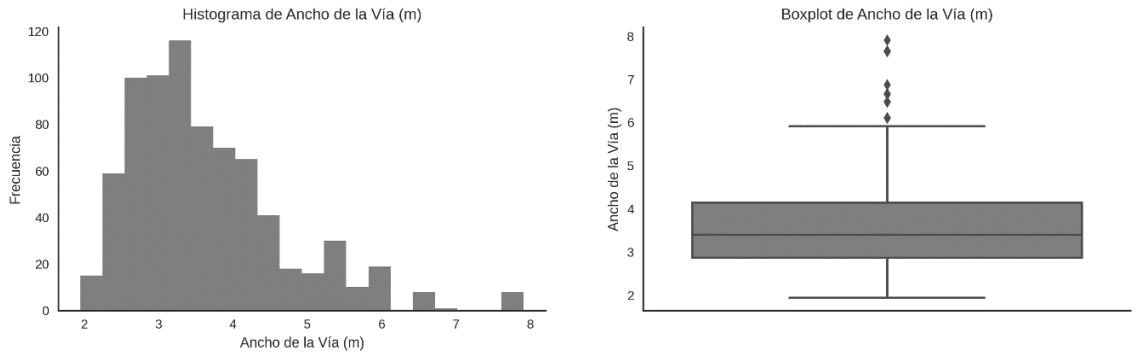
El ancho promedio de las vías es de 3.64 metros, mostrando una variabilidad significativa en su diseño. En cuanto a la longitud, las vías varían desde caminos muy cortos hasta tramos de casi 20 kilómetros, reflejando la diversidad de la red. La frecuencia de mantenimiento tiene un promedio de 13 meses, pero varía ampliamente, lo que sugiere diferentes necesidades a lo largo de la red.

La velocidad promedio en estas vías es de 23.32 km/h, aunque este número varía bastante, probablemente debido a las diferencias en las condiciones de las vías y el tráfico. El número de curvas por vía también muestra una amplia gama, lo que podría influir en la seguridad y el mantenimiento. Aunque la mayoría de las vías no tienen puentes, algunas tienen hasta cuatro.

El tráfico promedio diario anual en estas vías es relativamente bajo, pero hay una considerable variación que indica desde vías poco transitadas hasta algunas con un tráfico considerable. La producción económica asociada a las áreas de influencia de estas vías varía enormemente, reflejando la diversidad económica en la región. Del mismo modo, el número de viviendas y la población en estas áreas muestran variaciones significativas, lo que indica diferencias en la densidad poblacional.

En la ***figura 5*** se evidencia que la mayoría de los anchos de vía se concentran en un rango más bajo, aproximadamente entre 2 y 4 metros, pero hay una extensión hacia anchos mayores, lo que indica una asimetría hacia la derecha en la distribución. Así mismo los valores atípicos se presentan a partir de 6 metros, como se muestra en el gráfico.

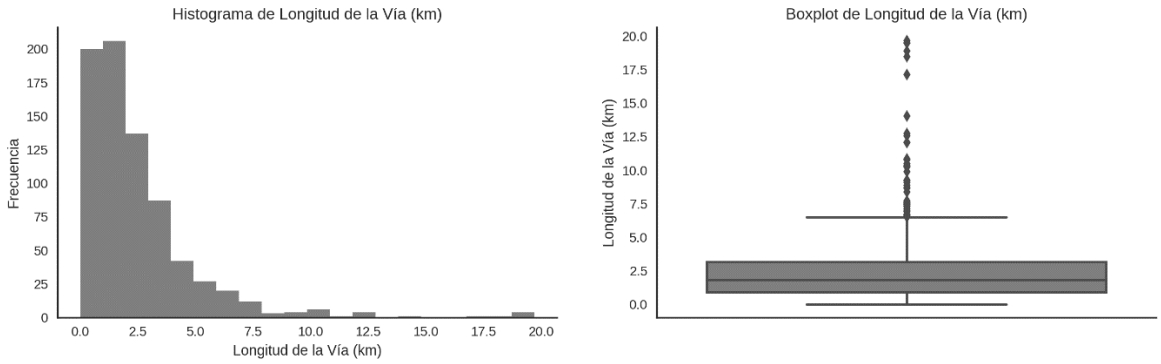
Figura 5. Histograma y Boxplot de Ancho de vía (m)



Fuente: Pérez. M, 2023

Las longitudes de vía se agrupan en un rango más bajo, principalmente entre 0 y 5 kilómetros, destacando la presencia de numerosos segmentos cortos en la red vial. Sin embargo, se observa una extensión hacia longitudes mayores, reflejando una asimetría en la distribución hacia el extremo superior. Los valores atípicos, representando vías inusualmente largas, se identifican en longitudes superiores a 10 kilómetros, como se muestra en la ***figura 6***.

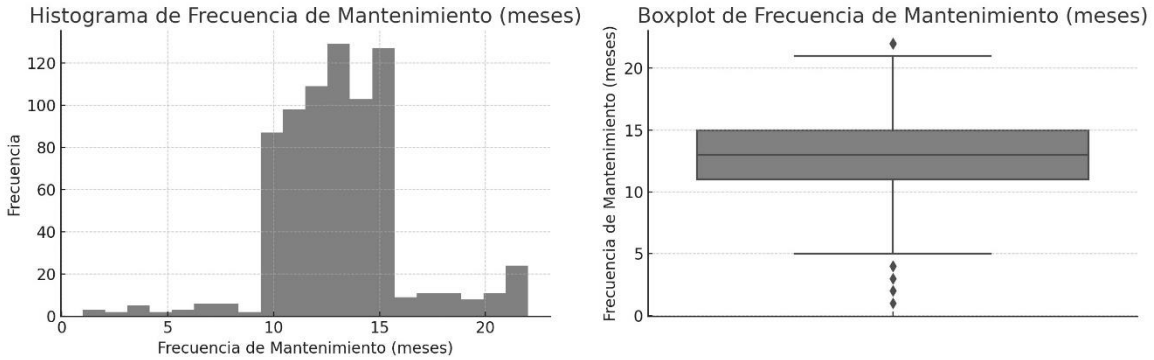
Figura 6. Histograma y Boxplot de Longitud de la vía (km)



Fuente: Pérez. M, 2023

La ***figura 7***, muestra que la frecuencia de mantenimiento vial se concentra principalmente entre 11 y 15 meses. La mediana, marcada en el *boxplot*, sugiere que la mitad de las vías reciben mantenimiento aproximadamente cada 13 meses. Los valores atípicos indican que algunas vías tienen frecuencias de mantenimiento significativamente diferentes, lo que puede señalar necesidades especiales o desafíos en ciertas áreas.

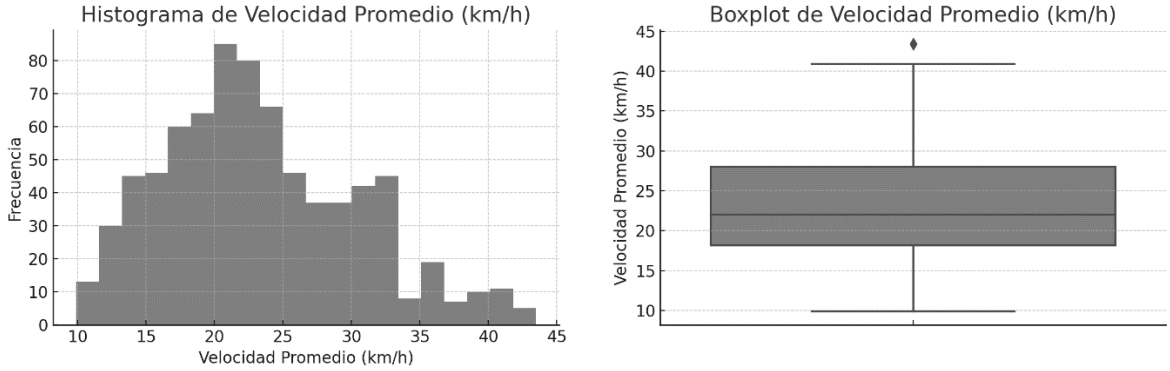
Figura 7. Histograma y Boxplot de Frecuencia de Mantenimiento (meses)



Fuente: Pérez. M, 2023

En la ***figura 8***, se observa, la distribución de la velocidad promedio en las vías rurales. El histograma muestra que la mayoría de las vías tienen velocidades promedio que oscilan entre aproximadamente 15 y 30 km/h, con el mayor pico alrededor de 20 km/h, indicando que es la velocidad más común. El *boxplot* refleja una mediana cercana a este rango y revela algunos valores atípicos que exceden los 35 km/h, sugiriendo que hay vías con velocidades significativamente superiores a la media.

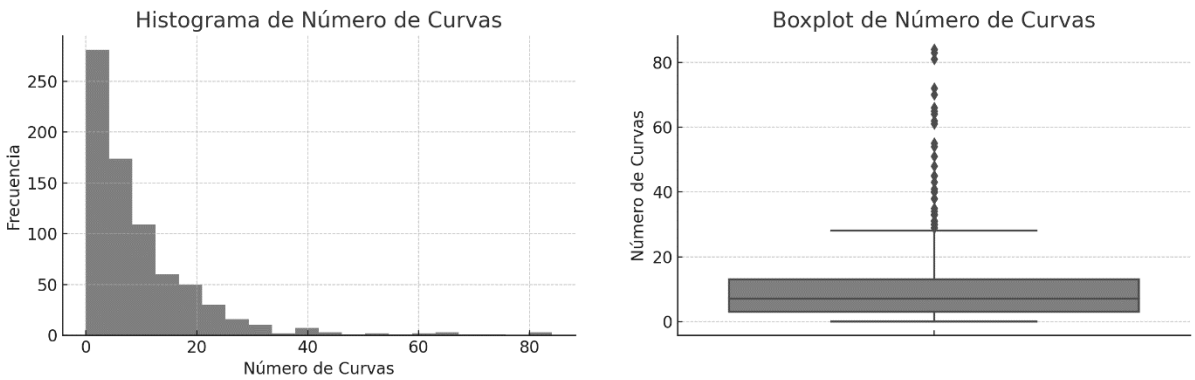
Figura 8. Histograma y Boxplot de Velocidad promedio (km/h)



Fuente: Pérez. M, 2023

La ***figura 9,*** presenta el número de curvas en las vías analizadas. En el histograma, se observa una concentración alta de vías con pocas curvas, disminuyendo la frecuencia a medida que aumenta el número de curvas. La mayoría de las vías tienen menos de 20 curvas, lo que sugiere que son relativamente rectas. El *boxplot* muestra una mediana baja y varios valores atípicos, lo que indica que hay vías con un número de curvas mucho mayor que el promedio.

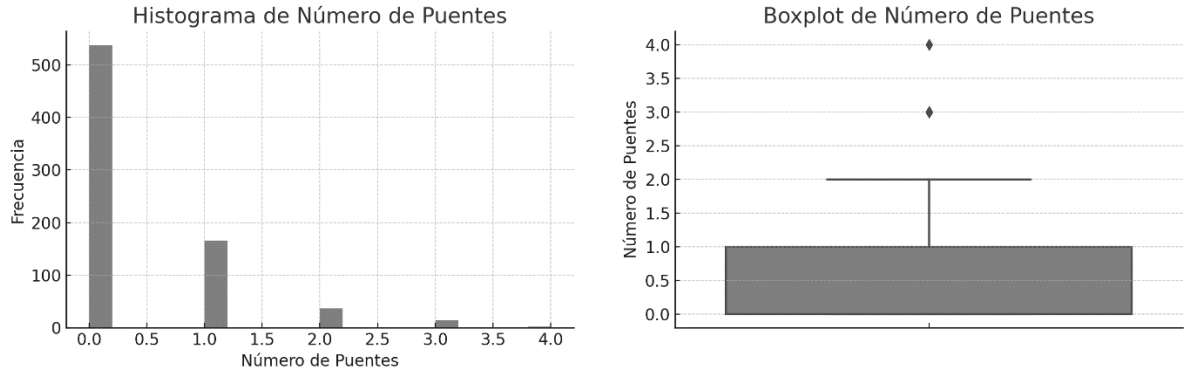
Figura 9. Histograma y Boxplot de Número de Curvas



Fuente: Pérez. M, 2023

En la ***figura 10***, la mayoría de las vías tienen pocos o ningún puente, como se observa en el pico prominente del histograma cerca de cero. El *boxplot* confirma esto, indicando que la mediana de puentes por vía es baja, y resalta algunos valores atípicos donde el número de puentes es más alto.

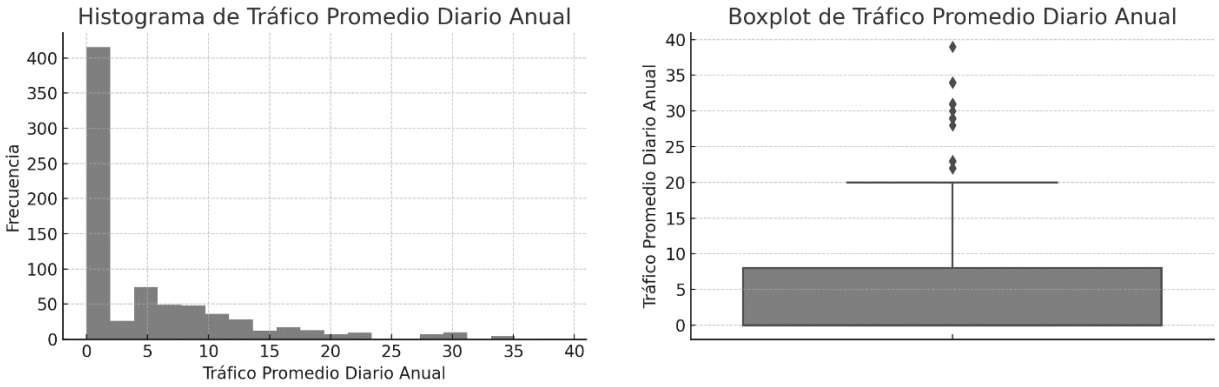
Figura 10. Histograma y Boxplot de Número de Puentes



Fuente: Pérez. M, 2023

Según muestra en la ***figura 11***, la mayoría de las vías tienen un tráfico bajo, con un pico significativo en el histograma cerca del valor cero. Esto indica que hay una gran cantidad de vías rurales con muy poco tráfico diario. El *boxplot* refleja que la mediana está cerca de los valores más bajos, lo que confirma que un tráfico reducido es lo más común. Sin embargo, hay varios valores atípicos que sugieren que algunas vías experimentan un tráfico mucho más alto de lo habitual.

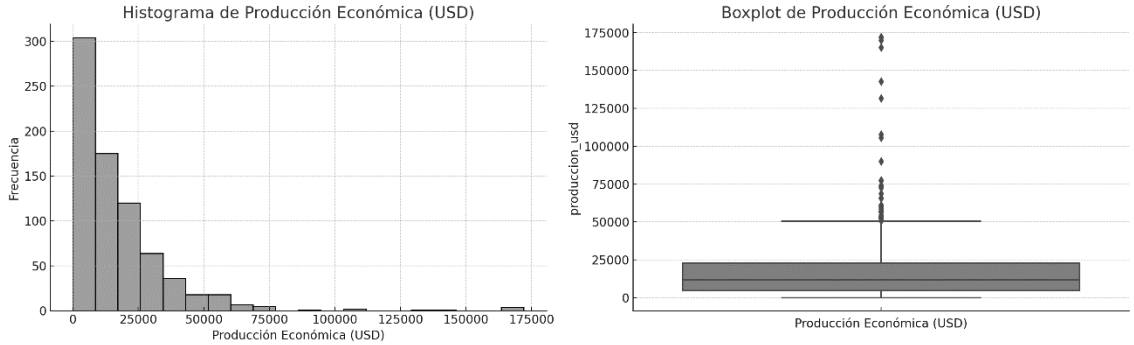
Figura 11. Histograma y Boxplot de Tráfico Promedio Diario Anual (TPDA)



Fuente: Pérez. M, 2023

La ***figura 12,*** muestra que la mayoría de las vías rurales tienen una producción económica relativamente baja, con un gran número de segmentos concentrados cerca de valores bajos en el histograma. Esto se refleja en el *boxplot*, donde la mediana está cerca del límite inferior y los 'bigotes' se extienden solo ligeramente, lo que sugiere que la mayor parte de la producción económica se mantiene dentro de un rango bajo a moderado. Sin embargo, los valores atípicos que se extienden hacia cifras de producción mucho más altas indican que algunos segmentos de la red vial tienen un impacto económico significativamente mayor, lo que podría justificar una inversión y mantenimiento prioritarios.

Figura 12. Histograma y Boxplot de Producción Económica (USD).

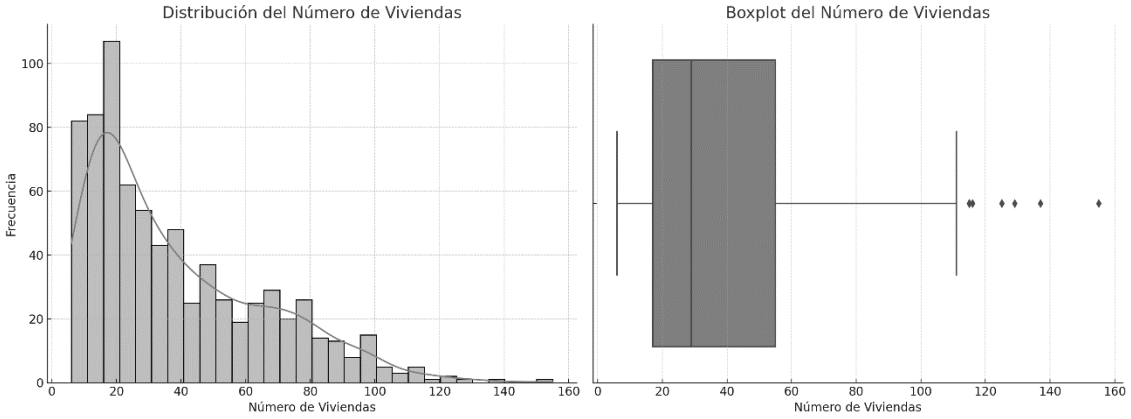


Fuente: Pérez. M, 2023

En la ***figura 13,*** se evidencia una gran concentración de tramos de carreteras con un número bajo de viviendas, indicando áreas rurales menos desarrolladas o con baja densidad de población. El *boxplot* revela que la mediana de viviendas por tramo de carretera es baja, reforzando la idea de que la mayoría de las carreteras transcurren por

zonas con pocas viviendas. Los valores atípicos sugieren la existencia de algunos tramos con una cantidad inusualmente alta de viviendas.

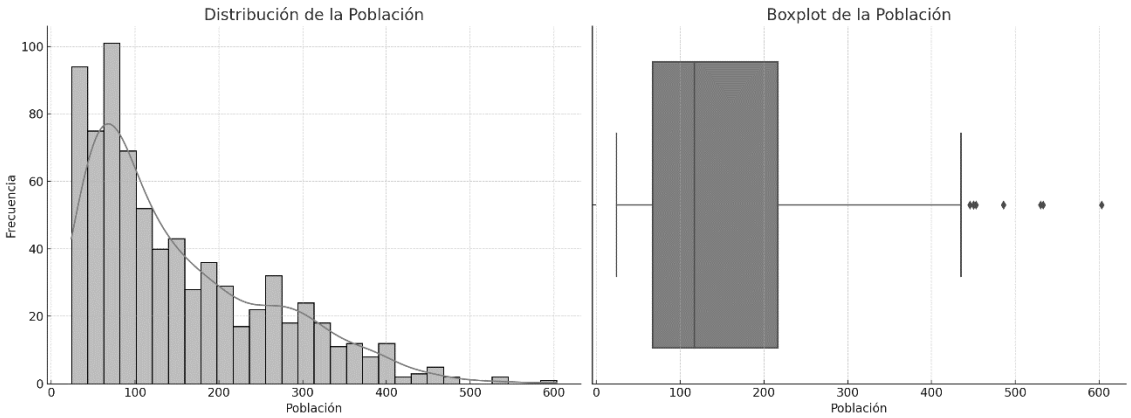
Figura 13. Histograma y Boxplot de Número de Viviendas



Fuente: Pérez. M, 2023

En la ***figura 14***. El histograma indica que la mayoría de las vías están en áreas con poblaciones pequeñas, como lo evidencia la concentración de la frecuencia hacia el extremo más bajo de la escala. El *boxplot* complementa esta observación mostrando una mediana baja y los cuartiles concentrados en el extremo inferior del rango de datos, lo cual confirma que una población reducida es común en la mayoría de las áreas. Por otro lado, los valores atípicos indican la presencia de vías que sirven a comunidades significativamente más grandes.

Figura 14. Histograma y Boxplot de Población



Fuente: Pérez. M, 2023

3.3 Preparación de los Datos

La preparación de datos es una etapa crítica en el proceso de análisis de datos, definida por (Pyle, 1999) como el proceso de limpiar y transformar los datos brutos antes

del análisis y la interpretación. Este proceso es esencial para garantizar la precisión, la integridad y la idoneidad de los datos para el análisis. Dasu y Johnson (Dasu, 2003) enfatizan que esta fase puede consumir entre el 50% y el 80% del tiempo total del proyecto, destacando su rol fundamental en la eficacia del análisis de datos.

3.3.1 Selección de Datos

Para el análisis en este proyecto, se emplearán todos los registros disponibles en las tablas que componen la base de datos. Dado que la base de datos ha sido específicamente diseñada y compilada para este proyecto, la cantidad y naturaleza de los registros insertados responden a una selección intencionada y estratégica.

No obstante, dentro de estos registros existen ciertos campos que no son pertinentes para los objetivos específicos de nuestro análisis de minería de datos. Por tanto, se realizará una cuidadosa selección de campos, excluyendo aquellos que no aportan valor significativo a nuestro estudio. Esta selección enfocada asegura la relevancia y eficiencia del análisis, al concentrarse en los datos más relevantes.

Figura 15. Datos seleccionados

```
dfviasm.info()
✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 756 entries, 0 to 755
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   codparr         756 non-null    int64
 1   clima           756 non-null    object
 2   c_terreno_tipo  756 non-null    int64
 3   longitud_km     756 non-null    float64
 4   ancho_mt        756 non-null    float64
 5   meses_manten    756 non-null    int64
 6   estado_via      756 non-null    int64
 7   c_rodea_via     756 non-null    int64
 8   carrilles       756 non-null    int64
 9   velocidad_prom  756 non-null    float64
 10  ncurvas         756 non-null    int64
 11  npuentes        756 non-null    int64
 12  tpda            756 non-null    int64
 13  produccion_usd  756 non-null    float64
 14  nviviendas      756 non-null    int64
 15  poblacion       756 non-null    int64
dtypes: float64(4), int64(11), object(1)
```

Fuente: Pérez. M, 2023

3.3.2 Limpieza de Datos

La base de datos con la que se cuenta para el proyecto contiene toda la información necesaria para poder cumplir los objetivos de la minería de datos, además, estos datos son

de dominio público y se determinan como datos reales, son datos relativamente limpios por lo que hace una limpieza rápida de los datos.

Verificamos la no existencia de valores nulos o duplicados:

Figura 16. Nulos y Duplicados

Fuente: Pérez. M, 2023

3.3.3 Construir Datos.

La variable clima se considera importante para el análisis en su forma actual es del tipo *object* por lo que es necesario convertirla a una variable de tipo numérica.

Figura 17. Conversión a Numérica de la Variable "Clima"

```
from sklearn.preprocessing import LabelEncoder

# Seleccionamos la variable y las categorias
dfviasm = pd.DataFrame({'clima': ['TROPICAL MEGATERMICO HUMEDO', 'TROPICAL MEGATERMICO SEMI HUMEDO']})

# Usar LabelEncoder para convertir en valores numéricos
encoder = LabelEncoder()
dfviasm['climat'] = encoder.fit_transform(dfviasm['clima'])

# Ahora, df contiene una columna adicional 'categoria_numerica' con valores numéricos
print(dfviasm)
```

✓ 0.0s

```
                          clima  climat
0       TROPICAL MEGATERMICO HUMEDO       0
1  TROPICAL MEGATERMICO SEMI HUMEDO       1
```

Fuente: Pérez. M, 2023

3.3.4 Integrar Datos.

No ha sido necesaria la creación de nuevas estructuras (campos, registros, etc.), ni la fusión entre distintas tablas de la base de datos.

3.3.5 Formato de los Datos.

El formado de los datos incluyendo la variable convertida se describen en la figura.

Figura 18. Formato de Dataset Vial

```
dfviasm.dtypes
```

✓ 0.0s

```
codparr             int64
clima               int64
c_terreno_tipo      int64
longitud_km       float64
ancho_mt          float64
meses_manten        int64
estado_via          int64
c_rodea_via         int64
carrilles           int64
velocidad_prom    float64
ncurvas             int64
npuentes            int64
tpda                int64
produccion_usd    float64
nviviendas          int64
poblacion           int64
dtype: object
```

Fuente: Pérez. M, 2023

3.4 Modelado

En esta fase de la metodología CRISP-DM, el modelado comienza con un análisis de correlación para seleccionar las variables más relevantes, seguido de un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos. Posteriormente, se aplicará el algoritmo de K-Means para agrupar los segmentos viales en clústeres basados en sus características similares. Este enfoque integrado nos permitirá identificar patrones y necesidades específicas de mantenimiento en la red vial, facilitando la creación de estrategias de mantenimiento más eficientes y focalizadas.

3.4.1 Selección de técnicas de modelado

Se seguirá un enfoque estructurado y metódico en la selección de técnicas de modelado. Inicialmente, se realizará un análisis de correlación para identificar y seleccionar las variables más influyentes. Este paso es fundamental para asegurar que solo se consideren las variables que aportan información significativa para el análisis.

Posteriormente, se aplicará un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos. PCA facilitará la identificación de las principales características que influyen en los segmentos viales, simplificando los datos sin perder información crítica.

Finalmente, el modelado culminará con la aplicación del algoritmo *K-means* para la creación de clústeres. Este método agrupará los segmentos viales en función de sus características similares derivadas del PCA, lo que permitirá una segmentación clara y significativa. Una vez formados los clústeres, se procederá a describir y analizar cada uno de ellos para obtener *insights* detallados sobre las necesidades específicas de mantenimiento en diferentes áreas de la red vial. Este enfoque integrado proporcionará una comprensión profunda y operativamente relevante para la planificación y ejecución del mantenimiento vial.

3.4.1.1 Escoger la técnica de Modelado

Se implementará un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos y resaltar las características principales de los segmentos viales. Finalmente, se utilizará el algoritmo *K-means* para la creación de clústeres basados en las características identificadas por el PCA. Este proceso culminará con un análisis detallado de cada clúster. La fórmula principal en *K-Means* involucra la minimización de la suma de cuadrados dentro de cada clúster:

Minimizar $\sum_{i=1}^{k} \sum_{x \in c_i} \|x - u_i\|^2$

Donde:

k es el número de clústeres.

Ci es el conjunto de puntos en el clúster ***i***.

x es un punto en el espacio de las características.

μi es el *centroide* del clúster ***i***.

El PCA transforma un conjunto de variables posiblemente correlacionadas en un conjunto de valores de variables no correlacionadas llamadas componentes principales. (Jolliffe I. T., 2002).

3.4.1.2 Generación de Modelos

En la generación de modelos se detallarán los parámetros y hiperparámetros empleados en el ajuste y modelado de las técnicas de análisis no supervisado seleccionadas. Siguiendo la preparación de datos previamente realizada, nuestro conjunto de datos incluye información crucial relacionada con aspectos de mantenimiento vial, como la longitud de los tramos, su estado actual, la velocidad promedio permitida, el número de curvas, el tráfico promedio diario anual (TPDA), y la producción económica asociada a cada segmento.

Para el Análisis de Componentes Principales (PCA), se ajustarán parámetros clave que permitan maximizar la varianza explicada con el menor número de componentes posibles. Esto facilitará la comprensión y visualización de las relaciones entre las variables.

Posteriormente, para la aplicación del algoritmo *K-means*, se definirán hiperparámetros como el número de clústeres, lo cual podría determinarse mediante técnicas como el método del codo o el análisis de silueta. Además, se establecerán los criterios para la asignación inicial de centroides y se definirán las iteraciones necesarias para la convergencia del algoritmo.

Este enfoque nos permitirá identificar patrones inherentes en los datos, agrupando segmentos viales con características similares y facilitando la toma de decisiones estratégicas para el mantenimiento vial eficiente.

3.5 Evaluación del Modelo

Para probar la calidad y validez de los clústeres, se empleará un enfoque adaptado de evaluación, dada la naturaleza no supervisada del análisis. Como métrica se utilizará *Silhouette Score*, que es utilizada para medir la calidad del agrupamiento de un algoritmo de *clustering*, como *K-Means*, indicando cuán bien separados se encuentran los grupos. El cálculo del *Silhouette Score* para cada punto de datos se basa en dos medidas:

1. **Cohesión (a):** La distancia media entre un punto y todos los otros puntos en el mismo clúster.
2. **Separación (b):** La distancia media entre un punto y todos los puntos en el clúster más cercano al que el punto no pertenece.

La fórmula del *Silhouette Score* para un solo punto es la siguiente:

$$\boldsymbol{S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}}$$

Donde ***S(i)*** es el *Silhouette Score* del punto ***i***, ***a(i)*** es la medida de cohesión y ***b(i)*** es la medida de separación para el punto ***i.***

- El valor de ***S(i)*** varía entre -1 y 1.
- Un valor alto indica que el punto está bien agrupado y lejos de los clústeres vecinos.
- Un valor cercano a 0 indica que el punto está cerca de la frontera de decisión entre dos clústeres vecinos.
- Un valor negativo indica que el punto podría haber sido asignado al clúster equivocado.

Para obtener un *Silhouette Score* general para el modelo, se calcula el promedio de los *Silhouette Scores* de todos los puntos de datos. Este promedio proporciona una medida de cuán apropiados son los clústeres formados para los datos. Un valor más alto indica una mejor definición de los clústeres (Rousseeuw, 1987).

Capítulo 4.

En este capítulo se aborda dos aspectos principales: la aplicación de análisis de clústeres generados para modelar datos viales y la utilización de *Power BI* para la visualización y generación del cronograma de mantenimiento vial. Se detalla la metodología para formar clústeres y posteriormente se describe cómo estos se presentan visualmente, proporcionando una herramienta dinámica para la planificación del mantenimiento vial.

4.1 Caracterización Clúster

Variables importantes:

- ***Longitud_Km:*** Extensión del tramo
- ***Estado_via:*** Estado de la vía, Bueno=1, regular=2 y malo=3
- ***Velocidad_prom:*** Velocidad promedio de circulación en el tramo
- ***Ncurvas***: Número de curvas en el tramo
- ***Tpda:*** TPDA (Tráfico Promedio Diario Anual)
- ***Produccion_usd:*** Producción estimada en dólares

4.1.1 Varianza Explicada.

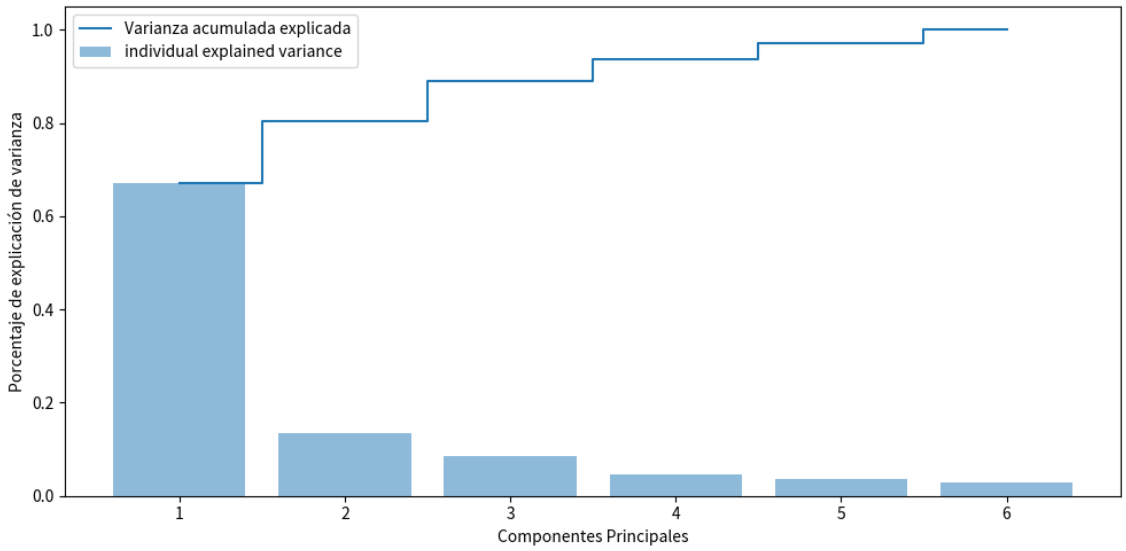
En el marco del análisis de componentes principales (PCA) presentado en la ***Tabla 5***, el primer componente principal explica el 67,1% de la varianza, siendo el más significativo. Los dos primeros componentes combinados explican el 80,5%, destacando su importancia en capturar la esencia de los datos. El tercer componente añade un 8,6%, llevando la varianza acumulada al 89,1%. Esto resalta que la mayor parte de la información se concentra en los tres primeros componentes, subrayando su relevancia en la representación comprensiva de los datos.

Tabla 5. Varianza Explicada según Componente

Componente	Varianza Explicada	Varianza Acumulada
1	67,1%	67,1%
2	13,4%	80,5%
3	8,6%	89,1%
4	4,6%	93,6%
5	3,5%	97,2%
6	2,8%	100,0%

Fuente: Pérez. M, 2024

Figura 19. Varianza Explicada según Componente



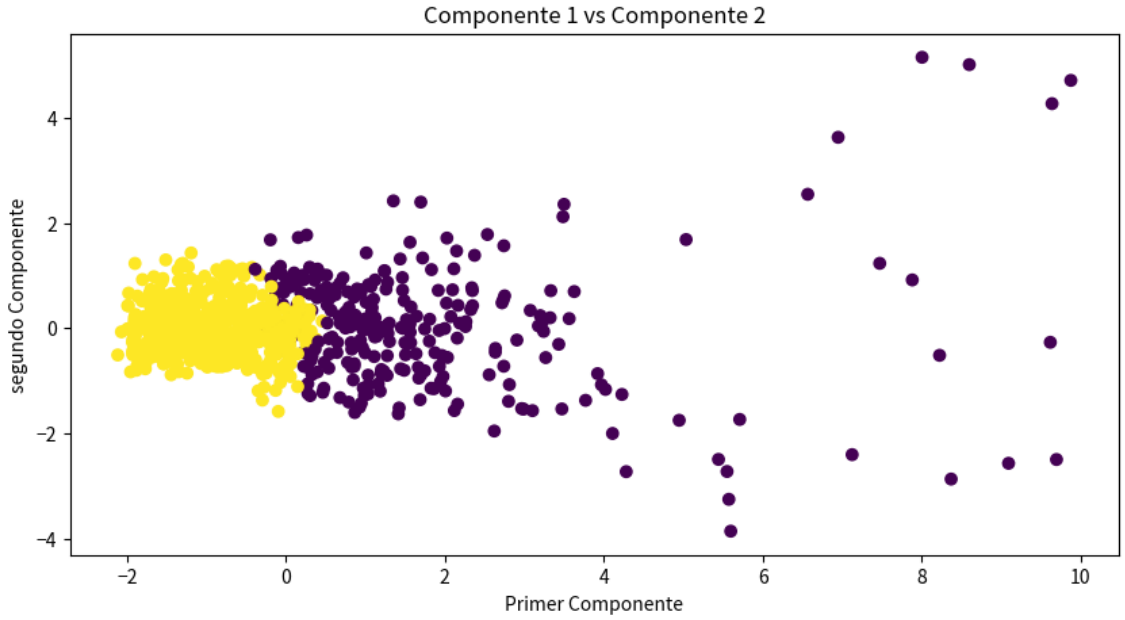
Fuente: Pérez. M, 2024

En consecuencia, para la construcción del Clúster se tomó en consideración los 3 primeros compontes que resumen el 89,1% de la varianza.

4.1.2 Clusterización: K-means.

Se obtuvo un clúster con dos segmentos y una representatividad del 72.9% con 617 casos para el clúster 0 y 138 para el clúster 1.

Figura 20. Dispersión Entre la Primera y Segunda Componente



Fuente: Pérez. M, 2024

4.1.3 Componentes de tamaño y forma

La efectividad del Análisis de Componentes Principales (PCA) para discernir las dimensiones subyacentes de un conjunto de datos se apoya en la interpretación precisa de sus componentes principales. Esta interpretación se fundamenta en las cargas de las variables, que son indicativas de la correlación entre las variables originales y cada componente principal. Según (Jolliffe I. T., 2016), el acceso a los vectores propios del PCA es esencial para identificar estas cargas, ya que cada vector propio, asociado a una componente principal específica, contiene los coeficientes (cargas) que reflejan el grado de influencia de cada variable original en la componente.

Para profundizar en la estructura de los datos, es primordial examinar las cargas de las variables en las primeras tres componentes principales. Esta exploración revela cómo las variables originales contribuyen a las dimensiones más significativas capturadas por el PCA, permitiendo una interpretación de los datos en términos de sus componentes de tamaño y forma.

(Jolliffe I. T., 2016) enfatiza que, aunque la interpretación de las componentes principales puede variar según el analista y el contexto específico del estudio, en general, las variables con altas cargas en una componente indican una contribución significativa a la varianza explicada por dicha componente.

Este proceso de interpretación, al ser aplicado a las dos primeras componentes principales, no solo facilita un entendimiento más profundo de las principales fuerzas que modelan el conjunto de datos, sino que también subraya la flexibilidad del PCA para adaptarse a diferentes tipos y estructuras de datos. La subjetividad inherente en la interpretación de las componentes principales resalta la importancia de un análisis contextualizado y consciente de las particularidades del conjunto de datos analizado.

Tabla 6. cargas de los Vectores Según Variables Implicadas

Variable	PC1	PC2
longitud_km	0,5934	0,1267
estado_via	-0,0051	-0,0636
velocidad_prom	0,0361	0,5852
ncurvas	0,5134	-0,6564
tapa	0,3366	0,0129
produccion_usd	0,5193	0,4545

Fuente: Pérez. M, 2024

En la ***Tabla 6***, se presenta un patrón de asociación diferenciado entre dos componentes principales y las variables examinadas. El primer componente principal muestra una fuerte relación con 'longitud_km', 'ncurvas' y 'produccion_usd'. En contraste, el segundo componente principal está vinculado estrechamente con 'velocidad_prom', 'ncurvas' y 'produccion_usd'. Estos hallazgos se pueden interpretar de la siguiente forma:

PC1 (Componente de Magnitud): El primer componente principal se caracteriza por altas cargas factoriales en variables como 'longitud_km', 'ncurvas' y 'produccion_usd', lo que indica que este componente podría interpretarse como un indicador de magnitud. Representa una dimensión que integra tanto la extensión física de los tramos de carretera (longitud y cantidad de curvas) como el impacto económico asociado a dicha extensión (producción en dólares), reflejando así la relevancia y el peso que estas variables tienen en conjunto.

PC2 (Componente de Diferenciación): El segundo componente principal se distingue por la coexistencia de cargas factoriales negativas en 'estado_via' y 'ncurvas', junto con una carga positiva significativa en 'velocidad_prom' y 'produccion_usd'. Esto sugiere que el PC2 captura un contraste en las características de los tramos carreteros, destacando aquellos que, a pesar de ser más cortos y contar con menos curvas, permiten una mayor velocidad promedio de tránsito y no por ello dejan de tener una contribución económica relevante. Este componente puede interpretarse como una medida de eficiencia o productividad de los tramos de carretera.

4.1.4 Caracterización de segmentos utilizando las variables principales

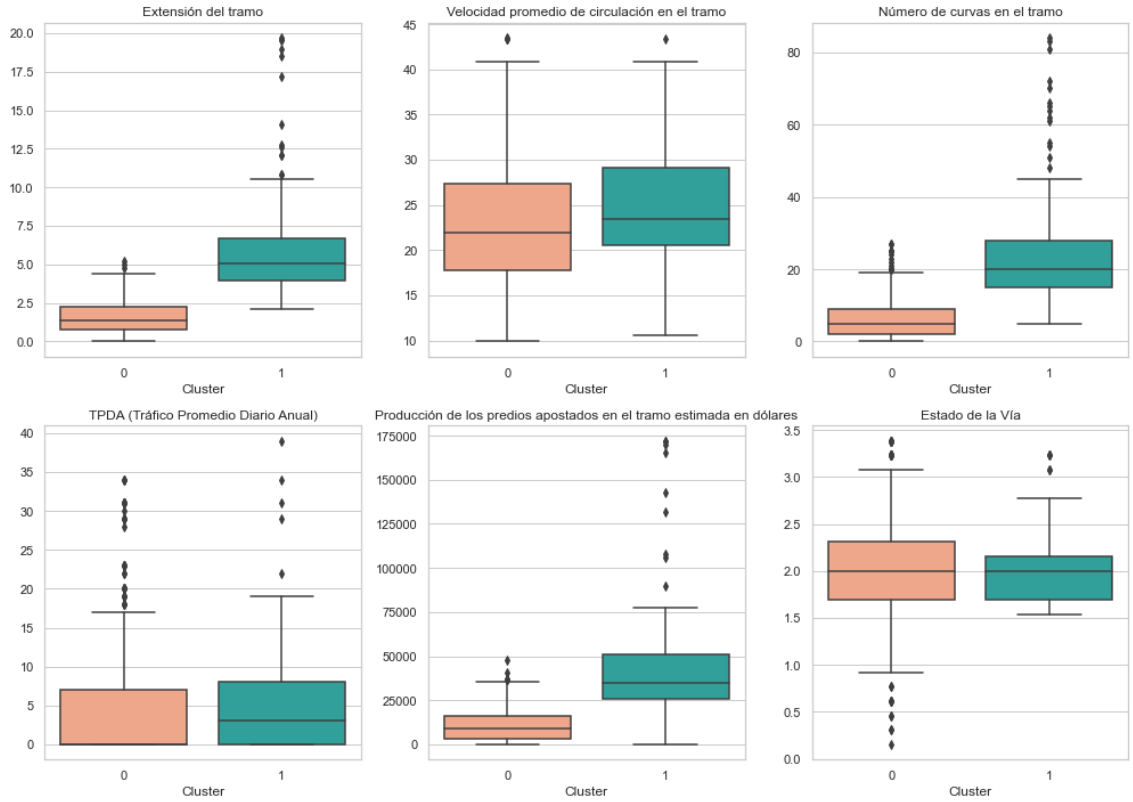
Para la caracterización de los segmentos se utilizarán las variables principales e implicadas en su construcción, esto con el fin de mostrar las diferencias significativas entre los dos grupos identificados. La selección de las variables de 'extensión del tramo', 'velocidad promedio de circulación', 'número de curvas', 'tráfico promedio diario anual (TPDA)', 'producción estimada de los predios', y 'estado de la vía', ha sido determinante para resaltar las características distintivas de cada conglomerado.

Esta diferenciación basada en variables clave permite una comprensión más profunda de las necesidades y prioridades de cada segmento, y establece una base sólida para el desarrollo de estrategias de intervención específicas y eficaces.

La **Figura 21** muestra en conjunto, que los resultados delinean un perfil para cada cluster: el *Cluster 0* está asociado con tramos más cortos, mayor fluidez de tráfico y mejor estado de la vía. Por contraste, el *Cluster 1* se define por tramos más extensos y sinuosos, con menor flujo vehicular y un estado general de la vía más deteriorado, aunque con una producción económica vinculada a los predios adyacentes considerablemente más alta.

Estas características son fundamentales para la comprensión y el posterior desarrollo de estrategias de gestión vial y planeación territorial.

Figura 21. Bloxplot de Variables Principales



Fuente: Pérez. M, 2024

Al analizar los datos para la construcción de clústeres, se realizaron dos aproximaciones para el tratamiento de *outliers*: una consideración directa de los datos y una segunda aproximación aplicando el rango intercuartílico. Los resultados de ambas aproximaciones permiten contrastar el impacto del tratamiento de *outliers* en las variables estudiadas.

Valores promedio de variables utilizadas para construir el clúster

Tabla 7. Análisis Preliminar de Clústeres sin Tratamiento de Outliers

Clúster	Longitud_km	Estado_via	Velocidad_prom	Ncurvas	Tpda	Produccion_usd
0	1,62	2,08	23,05	6,28	5,20	$ 10.986,70
1	6,33	2,09	24,51	25,89	10,37	$ 44.186,80

Fuente: Pérez. M, 2024

Tabla 8. Análisis de Clústeres Post-Tratamiento de Outliers con Rango Intercuartílico

Clúster	Longitud_km	Estado_via	Velocidad_prom	Ncurvas	Tpda	Produccion_usd
0	1,63	2,00	23,04	6,29	5,21	$ 11.017,65
1	5,19	2,00	24,55	20,89	10,34	$ 36.322,39

Fuente: Pérez. M, 2024

Para el **Clúster 0**, antes del tratamiento de *outliers*, se observa que la longitud del tramo varía entre 2.13 km y 19.69 km, con una media de 6.33 km. Tras aplicar el rango intercuartílico, la longitud promedio se ajusta a 1.63 km, indicando una disminución en la variabilidad de la longitud de los tramos considerados, posiblemente por la eliminación de valores extremos. De manera similar, el estado de la vía, la velocidad promedio, el número de curvas, el TPDA, y la producción en dólares experimentan ajustes leves, reflejando una tendencia hacia valores más representativos de la distribución central.

En cuanto al **Clúster 1**, después del tratamiento de *outliers*, se nota un cambio más significativo. La longitud promedio se ajusta de 6.33 km a 5.19 km, y el número de curvas promedio se reduce de 25.89 a 20.89, señalando una influencia notable de valores extremos en las mediciones originales. Las demás variables, como el estado de la vía, la velocidad promedio, el TPDA y la producción en dólares, también muestran ajustes, sugiriendo una caracterización más precisa de los tramos de carretera al excluir *outliers*.

Comparación entre Clústeres:

Inicialmente, se interpretaba que el Clúster 0 representaba tramos de carretera más cortos, con menos curvas, menor TPDA y menor producción en dólares en comparación con el Clúster 1. Sin embargo, tras el tratamiento de *outliers*, ambos clústeres exhiben cambios en sus valores promedio, aunque conservando patrones similares. El Clúster 1 continúa representando tramos de carretera más largos, con más curvas, mayor TPDA y

mayor producción en dólares. A pesar de los ajustes, la relación comparativa entre los clústeres se mantiene, lo que sugiere una segmentación robusta.

Implicaciones del Tratamiento de Outliers:

La aplicación del rango intercuartílico para el tratamiento de *outliers* ofrece una visión más detallada y representativa de las características de los tramos de carretera analizados. Al eliminar valores extremos, se logra una comprensión más enfocada de las condiciones promedio de las vías. Este proceso resalta la importancia de los métodos de tratamiento de datos para alcanzar conclusiones más fiables y representativas en el análisis de clústeres.

4.1.5 Etiquetas de los clústeres.

Clúster 0 - "Tramos de Atención Regular": Este clúster representa tramos de carretera que son más cortos, tienen un menor número de curvas y un TPDA más bajo. Además, estos tramos tienen una menor producción económica. Estas características sugieren que estos tramos de carretera son más sencillos y menos productivos.

Clúster 1 - "Tramos de Atención Prioritaria": Este clúster representa tramos de carretera que son más largos, tienen un mayor número de curvas y un alto Tráfico Promedio Diario Anual (TPDA). Además, *estos* tramos tienen una mayor producción económica. Estas características sugieren que estos tramos de carretera son más complejos y productivos.

4.1.6 Cruces con variables relevantes.

La **Tabla 9** resume la clasificación de los tramos viales en dos cantones, según la prioridad de mantenimiento. En *La Concordia*, de un total de 84 tramos, 13 son considerados de Atención Prioritaria y 71 de Atención Regular, mientras que, en *Santo Domingo*, de 672 tramos, 125 requieren atención prioritaria y 547 regular.

Tabla 9. Clúster Caracterizado según Cantón

Cantón	Tramos de Atención Prioritaria	Tramos de atención regular	Total
LA CONCORDIA	13	71	84
SANTO DOMINGO	125	547	672
Total	**138**	**618**	**756**

Fuente: Pérez. M, 2024

En la **Tabla 10** se evidencia que Alluriquín presenta el mayor número de Tramos de Atención Prioritaria con 32, seguido de Santo Domingo con 37, mientras que La Villegas registra el menor con solo 1. En cuanto a Tramos de Atención Regular, Puerto Limón lidera con 93, y Santo Domingo sigue con 166.

Tabla 10. Clúster Caracterizado según Parroquia

Parroquia	Tramos de Atención Prioritaria	Tramos de atención regular	Total
ALLURIQUIN	32	45	77
EL ESFUERZO	11	37	48
LA CONCORDIA	4	16	20
LA VILLEGAS	1	20	21
LUZ DE AMERICA	3	55	58
MONTERREY	6	24	30
PLAN PILOTO	2	11	13
PUERTO LIMON	9	93	102
SAN JACINTO	12	64	76
SANTA MARIA	10	22	32
SANTO DOMINGO	37	166	203
VALLE HERMOSO	11	65	76
Total	**138**	**618**	**756**

Fuente: Pérez. M, 2024

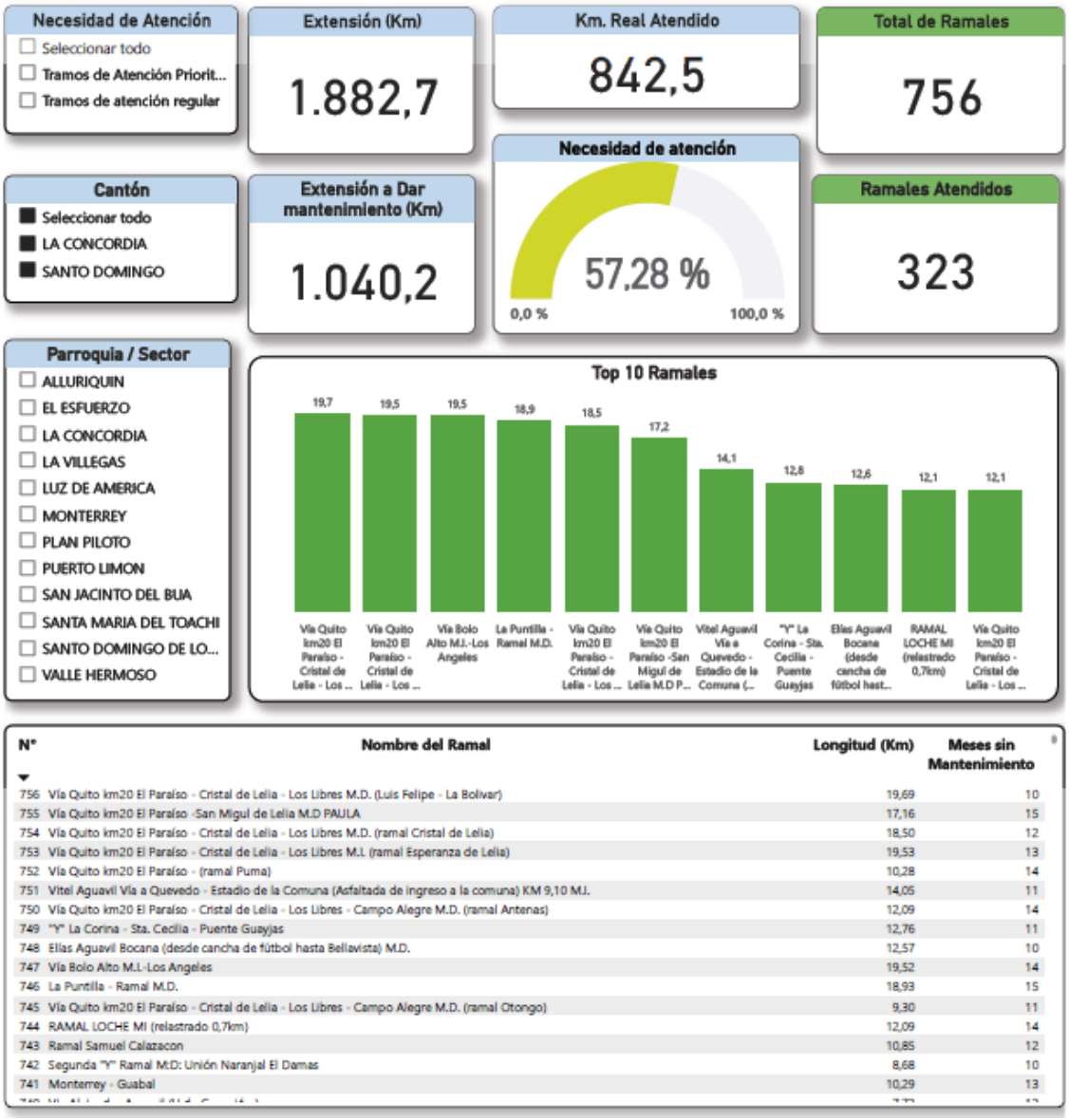
La suma total de tramos analizados asciende a 756, de los cuales 138 requieren atención prioritaria y 618 atención regular. Esta distribución resalta las áreas que demandan una atención urgente y las que precisan mantenimiento rutinario, subrayando la importancia de una gestión diferenciada y focalizada del mantenimiento vial en función de las necesidades específicas de cada parroquia.

4.2 Visualización.

La visualización mediante el *dashboard* de *Power BI* que muestra en la **Figura 22**, pretende optimizar la gestión del mantenimiento vial al categorizar los tramos en atención prioritaria o regular basado en la segmentación de datos previamente analizada. Incorpora variables claves como la extensión en kilómetros de los ramales, distinguiendo entre los

kilómetros pendientes de intervención y aquellos ya mantenidos. Facilita la selección de áreas específicas mediante filtros por cantón y parroquia y resalta los 10 ramales más críticos, mostrando su longitud y el tiempo sin mantenimiento. Este enfoque estratégico prioriza las intervenciones necesarias, asegurando una planificación y ejecución efectiva del mantenimiento vial.

Figura 22. Dashboard para determinación de cronograma de atención



N°	Nombre del Ramal	Longitud (Km)	Meses sin Mantenimiento
756	Vía Quito km20 El Paraíso - Cristal de Lelia - Los Libres M.D. (Luis Felipe - La Bolivar)	19,69	10
755	Vía Quito km20 El Paraíso -San Miguil de Lelia M.D PAULA	17,16	15
754	Vía Quito km20 El Paraíso - Cristal de Lelia - Los Libres M.D. (ramal Cristal de Lelia)	18,50	12
753	Vía Quito km20 El Paraíso - Cristal de Lelia - Los Libres M.L (ramal Esperanza de Lelia)	19,53	13
752	Vía Quito km20 El Paraíso - (ramal Puma)	10,28	14
751	Vitel Aguavil Vía a Quevedo - Estadio de la Comuna (Asfaltada de ingreso a la comuna) KM 9,10 M.L.	14,05	11
750	Vía Quito km20 El Paraíso - Cristal de Lelia - Los Libres - Campo Alegre M.D. (ramal Antenas)	12,09	14
749	"Y" La Corina - Sta. Cecilia - Puente Guayjas	12,76	11
748	Elías Aguavil Bocana (desde cancha de fútbol hasta Bellavista) M.D.	12,57	10
747	Vía Bolo Alto M.L-Los Angeles	19,52	14
746	La Puntilla - Ramal M.D.	18,93	15
745	Vía Quito km20 El Paraíso - Cristal de Lelia - Los Libres - Campo Alegre M.D. (ramal Otongo)	9,30	11
744	RAMAL LOCHE MI (relastrado 0,7km)	12,09	14
743	Ramal Samuel Calazacon	10,85	12
742	Segunda "Y" Ramal M.D: Unión Naranjal El Damas	8,68	10
741	Monterrey - Guabal	10,29	13

Fuente: Pérez. M, 2024, https://n9.cl/vializa

Esta herramienta resume el análisis y segmentación detallada de la base de datos vial, destacando los segmentos críticos para la intervención. Funciona como un puente entre la investigación y la aplicación práctica, traduciendo complejidades analíticas en *insights* accionables para la planificación y organización de mantenimientos viales. Al integrar

datos de segmentación y características viales, proporciona una línea base para decisiones informadas, asegurando una gestión eficiente y focalizada de los recursos en mantenimiento vial.

Capítulo 5. Conclusiones y Recomendaciones

Conclusiones

- La aplicación de algoritmos de aprendizaje no supervisado, como el Análisis de Componentes Principales (PCA) y el método K-means para la clusterización, ha revelado de manera efectiva la presencia de dos clústeres distintos en una red vial de 756 segmentos estudiados. La estratificación resultante en dos grupos ha facilitado la identificación de patrones y necesidades de mantenimiento específicos, permitiendo desarrollar una estrategia de gestión de mantenimiento vial orientada por los datos. Esta estrategia asigna los recursos de forma eficiente hacia las áreas más críticas y mejora los esfuerzos de conservación. La validación del modelo mediante el *Silhouette Score* ha corroborado la precisión de este enfoque, confirmando la adecuada diferenciación y cohesión entre los clústeres formados.
- El total de tramos viales es de 756, donde; el Clúster 0 con 638 casos, categorizado como "Tramos de Atención Regular", incluye tramos más cortos, con mejor fluidez y estado de la vía, indicando una menor complejidad y producción económica, y sugiriendo la necesidad de un mantenimiento regular que mantenga su condición actual. Contrariamente, el Clúster 1 con 138 casos, se describe como "Tramos de Atención Prioritaria", se caracteriza por tramos más extensos y sinuosos, con menor flujo vehicular, pero con un estado más deteriorado y una alta producción económica en los predios adyacentes, lo que resalta su importancia estratégica y la urgencia de un mantenimiento que mejore su infraestructura.
- Se destaca la potencialidad que tienen las técnicas de ML, particularmente el aprendizaje no supervisado, para revolucionar la planificación y gestión del mantenimiento vial. Al introducir métodos de caracterización que asignan prioridades de manera precisa y objetiva a cada vía, se superan las limitaciones anteriores, donde esta discriminación no era factible. Esta innovación no solo optimiza los recursos disponibles, sino que también incrementa notablemente la eficacia y eficiencia en la ejecución de tareas de mantenimiento vial, abriendo un camino hacia una gestión vial más inteligente y fundamentada en datos.
- El *dashboard* aporta una perspectiva valiosa sobre la gestión del mantenimiento vial, ofreciendo un claro panorama de la situación actual a través de los datos más relevantes. Los indicadores visuales y las estadísticas proporcionan una base para una

asignación de recursos más informada, destacando no solo las necesidades de mantenimiento sino también los avances logrados. Este enfoque basado en datos apoya una planificación estratégica dirigida a maximizar la eficiencia operativa y el bienestar de la infraestructura vial.

- Los 2 cantones de la provincia, incluyendo sus parroquias, exhiben tanto tramos de atención prioritaria como regular. A pesar de que La Concordia es más pequeño en comparación con Santo Domingo y cuenta con menos tramos viales, precisamente 84, presenta una proporción considerablemente alta de necesidad de atención, alcanzando un 71%. En contraste, la necesidad de atención a nivel provincial se estima alrededor del 57%. Este contraste resalta la importancia de evaluar y atender las necesidades de mantenimiento vial no solo por la cantidad de tramos, sino por la urgencia y el impacto en la comunidad.

Recomendaciones

- Los algoritmos de aprendizaje no supervisado, como PCA y K-means, han demostrado ser efectivos en la identificación de patrones y necesidades de mantenimiento específicos. Por lo tanto, se recomienda continuar utilizando estos algoritmos y explorar otros métodos de aprendizaje no supervisado que puedan ofrecer información adicional en la construcción e identificación de particularidades y patrones en los tramos viales, no solo en Santo Domingo, sino también en cualquier provincia o región
- Implementar estrategias de mantenimiento diferenciadas para cada clúster, ya que cada clúster tiene características y necesidades únicas, se recomienda implementar estrategias de mantenimiento diferenciadas para cada uno. Por ejemplo, el Clúster 0 podría beneficiarse de un mantenimiento regular para mantener su condición actual, mientras que el Clúster 1 podría requerir un mantenimiento más intensivo para mejorar su infraestructura.
- Los análisis de datos avanzados pueden revolucionar la planificación y gestión del mantenimiento vial. Se recomienda incorporar estos análisis en todas las etapas de la planificación y gestión del mantenimiento vial para optimizar los recursos disponibles y aumentar la eficacia y eficiencia en la ejecución de tareas de mantenimiento vial.
- El *dashboard* proporciona una visión clara de la situación actual y puede ser una herramienta valiosa para la toma de decisiones informada. Se recomienda actualizar

la información disponible en el panel de control, ya que esto permite asignar recursos de manera más informada y planificar estratégicamente para maximizar la eficiencia operativa y el bienestar de la infraestructura vial.

- Considerar la urgencia y el impacto en la comunidad al evaluar y atender las necesidades de mantenimiento vial, ya que no todas las vías son iguales y algunas pueden requerir atención más urgente que otras. Se recomienda considerar la urgencia y el impacto en la comunidad al evaluar y atender las necesidades de mantenimiento vial. Por ejemplo, aunque La Concordia tiene menos tramos viales que Santo Domingo, tiene una proporción alta de necesidad de atención, por lo que podría ser prioritario atender estas necesidades.

a. Bibliografía

(s.f.).

Agrawal, R. &. (1994). Fast algorithms for mining association rules. *In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, (Vol. 1215)*, 487-499.

Alpaydin, E. (2020). *Introduction to Machine Learning.* MIT Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Brynjolfsson, E. H. (2011). Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance? *SSRN Electronic Journal. DOI: 10.2139/ssrn.1819486*.

Bughin, J. C. (2017). *McKinsey Global Institute*. Obtenido de https://www.mckinsey.com/~/media/mckinsey/industries/advanced%20electronics/our%20insights/how%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/mgi-artificial-intelligence-discussion-paper.ashx

Chapman, P. C. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS Inc.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 21-26.

Dasu, T. &. (2003). *Exploratory Data Mining and Data Cleaning.* John Wiley & Sons.

Davenport, T. H. (2012). Obtenido de Harvard Business Review.: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Fayyad, U. P.-S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17(3)*, 37.

García, S. L. (2015). *Data Preprocessing in Data Mining. .* Springer.

Haas, R. &. (2003). *Pavement Management Systems.* McGraw-Hill Professional.

Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics.

INEC, I. N. (10 de 10 de 2023). *Visualizados del Censo 2022*. Obtenido de https://www.censoecuador.gob.ec/data-y-resultados/#pix-tab-398c8f9c-4977318

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31(8)*, 651-666.

James, G. W. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

James, G. W. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

Jolliffe, I. T. (2002). *Principal Component Analysis.* (Springer, Ed.) New York, NY.: Springer Series in Statistics.

Jolliffe, I. T. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374.

Jordan, M. I. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349(6245)*, 255-260.

Kaplan, A. &. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62(1)*, 15-25.

Koh, H. C. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, 64-72.

López García, J. R. (2017). *Modelización de la probabilidad de accidente laboral en función de las condiciones de trabajo mediante técnicas "Machine Learning".*

Loukides, M. (2010). *O'Reilly Media.* Obtenido de https://www.oreilly.com/radar/what-is-data-science/

Luger, G. F. (2004). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* Addison-Wesley.

Manyika, J. C. (2011). *McKinsey Global Institute.* Obtenido de https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Mayer-Schönberger, V. &. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think2013.* Houghton Mifflin Harcourt.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* The MIT Press.

Nina, J. (2022). *academia.edu.* Obtenido de Implementación del machine learning en el mantenimiento predictivo: https://www.academia.edu/82841788/Implementaci%C3%B3n_del_machine_learning_en_el_mantenimiento_predictivo?source=swp_share

PDyOT, G. P. (12 de 09 de 2020). *Portal Web GADPSDT*. Obtenido de http://gptsachila.gob.ec/index.php/la-provincia/pdot

PIARC. (2016). *Manual de Mantenimiento de Carreteras.* Asociación Mundial de la Carretera (PIARC).

Pieplow, D. A. (1995). *Gravel Roads: Maintenance and Design Manual. .* U.S. Department of Agriculture, Forest Service.

Provost, F. &. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* O'Reilly Media, Inc.

Pyle, D. (1999). *Data Preparation for Data Mining.* Morgan Kaufmann Publishers.

Pyle, D. (1999). *Data Preparation for Data Mining.* Morgan Kaufmann Publishers.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Roweis, S. T. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, 290(5500)*, 2323-2326.

Russell, S. &. (2016). *Artificial Intelligence: A Modern Approach.* Pearson.

Saltz, J. S. (2017). The Data Science Process: A Systematic Approach to Extracting Insight from Data. *Big Data Research, 5(2)*, 8-12.

Shalaby, A. &. (2004). Application of Data Mining Techniques for Pavement Maintenance Management. . *Journal of Infrastructure Systems, 10(4)*, 167-175.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5(4)*, 13-22.

SIL, G. P. (10 de 10 de 2023). *SIL*. Obtenido de Componente Vial: https://www.gptsachila.gob.ec/sil_gad/index.php/menu-componentes/submenu-territorial-vial?view=article&id=117

SIL, S. d. (2020). *Sistema de Información Local*. Obtenido de Componente Territorial: https://sil.gptsachila.gob.ec/index.php/menu-componentes/submenu-territorial-vial?view=article&id=117

Smith, K. L. (2001). *Design and Construction of Asphalt Paving Materials with Crumb Rubber Modifier. .* Transportation Research Record.

Sun, L. &. (2016). *Developing Predictive Models for Pavement Maintenance: A Machine Learning Approach.* Transportation Research Record.

Thompson, M. R. (1979). Road Surface Texture and Skid Resistance. *Transportation Engineering Journal of ASCE*.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *ournal of Economic Perspectives, 28(2)*, 3-28.

Wickham, H. &. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media, Inc.

Wirth, R. &. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*

Xie, P. D. (2016). On the Generalization of Denoising Autoencoders. *IEEE Transactions on Neural Networks and Learning Systems, 27(8)*, 1627-1639.

Zhao, D. &. (2002). Predictive Model for Pavement Condition Using Machine Learning. *Journal of Computing in Civil Engineering, 16(3)*, 200-209.

b. Trabajos citados

(s.f.).

Agrawal, R. &. (1994). Fast algorithms for mining association rules. *In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, (Vol. 1215)*, 487-499.

Alpaydin, E. (2020). *Introduction to Machine Learning.* MIT Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Brynjolfsson, E. H. (2011). Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance? *SSRN Electronic Journal. DOI: 10.2139/ssrn.1819486*.

Bughin, J. C. (2017). *McKinsey Global Institute*. Obtenido de https://www.mckinsey.com/~/media/mckinsey/industries/advanced%20electronics/our%20insights/how%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/mgi-artificial-intelligence-discussion-paper.ashx

Chapman, P. C. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS Inc.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 21-26.

Dasu, T. &. (2003). *Exploratory Data Mining and Data Cleaning.* John Wiley & Sons.

Davenport, T. H. (2012). Obtenido de Harvard Business Review.: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Fayyad, U. P.-S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17(3)*, 37.

García, S. L. (2015). *Data Preprocessing in Data Mining. .* Springer.

Haas, R. &. (2003). *Pavement Management Systems.* McGraw-Hill Professional.

Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics.

INEC, I. N. (10 de 10 de 2023). *Visualizados del Censo 2022*. Obtenido de https://www.censoecuador.gob.ec/data-y-resultados/#pix-tab-398c8f9c-4977318

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31(8)*, 651-666.

James, G. W. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

James, G. W. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

Jolliffe, I. T. (2002). *Principal Component Analysis.* (Springer, Ed.) New York, NY.: Springer Series in Statistics.

Jolliffe, I. T. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374.

Jordan, M. I. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349(6245)*, 255-260.

Kaplan, A. &. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62(1)*, 15-25.

Koh, H. C. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, 64-72.

López García, J. R. (2017). *Modelización de la probabilidad de accidente laboral en función de las condiciones de trabajo mediante técnicas "Machine Learning".*

Loukides, M. (2010). *O'Reilly Media.* Obtenido de https://www.oreilly.com/radar/what-is-data-science/

Luger, G. F. (2004). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* Addison-Wesley.

Manyika, J. C. (2011). *McKinsey Global Institute.* Obtenido de https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Mayer-Schönberger, V. &. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think2013.* Houghton Mifflin Harcourt.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* The MIT Press.

Nina, J. (2022). *academia.edu.* Obtenido de Implementación del machine learning en el mantenimiento predictivo: https://www.academia.edu/82841788/Implementaci%C3%B3n_del_machine_learning_en_el_mantenimiento_predictivo?source=swp_share

PDyOT, G. P. (12 de 09 de 2020). *Portal Web GADPSDT*. Obtenido de http://gptsachila.gob.ec/index.php/la-provincia/pdot

PIARC. (2016). *Manual de Mantenimiento de Carreteras.* Asociación Mundial de la Carretera (PIARC).

Pieplow, D. A. (1995). *Gravel Roads: Maintenance and Design Manual. .* U.S. Department of Agriculture, Forest Service.

Provost, F. &. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* O'Reilly Media, Inc.

Pyle, D. (1999). *Data Preparation for Data Mining.* Morgan Kaufmann Publishers.

Pyle, D. (1999). *Data Preparation for Data Mining.* Morgan Kaufmann Publishers.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Roweis, S. T. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, 290(5500)*, 2323-2326.

Russell, S. &. (2016). *Artificial Intelligence: A Modern Approach.* Pearson.

Saltz, J. S. (2017). The Data Science Process: A Systematic Approach to Extracting Insight from Data. *Big Data Research, 5(2)*, 8-12.

Shalaby, A. &. (2004). Application of Data Mining Techniques for Pavement Maintenance Management. . *Journal of Infrastructure Systems, 10(4)*, 167-175.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5(4)*, 13-22.

SIL, G. P. (10 de 10 de 2023). *SIL*. Obtenido de Componente Vial: https://www.gptsachila.gob.ec/sil_gad/index.php/menu-componentes/submenu-territorial-vial?view=article&id=117

SIL, S. d. (2020). *Sistema de Información Local*. Obtenido de Componente Territorial: https://sil.gptsachila.gob.ec/index.php/menu-componentes/submenu-territorial-vial?view=article&id=117

Smith, K. L. (2001). *Design and Construction of Asphalt Paving Materials with Crumb Rubber Modifier. .* Transportation Research Record.

Sun, L. &. (2016). *Developing Predictive Models for Pavement Maintenance: A Machine Learning Approach.* Transportation Research Record.

Thompson, M. R. (1979). Road Surface Texture and Skid Resistance. *Transportation Engineering Journal of ASCE*.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *ournal of Economic Perspectives, 28(2)*, 3-28.

Wickham, H. &. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media, Inc.

Wirth, R. &. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*

Xie, P. D. (2016). On the Generalization of Denoising Autoencoders. *IEEE Transactions on Neural Networks and Learning Systems, 27(8)*, 1627-1639.

Zhao, D. &. (2002). Predictive Model for Pavement Condition Using Machine Learning. *Journal of Computing in Civil Engineering, 16(3)*, 200-209.