ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

MAESTRÍA EN SISTEMAS DE INFORMACIÓN
MENCIÓN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

DISEÑO DE UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA GESTIÓN DE LA FLOTA VEHICULAR DE UNA EMPRESA PÚBLICA

PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN SISTEMAS DE INFORMACIÓN CON MENCIÓN EN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

CRISTIAN ROBERTO BENALCÁZAR DE LA CRUZ
crbdlc1@hotmail.com

DIRECTORA: PHD. TANIA ELIZABETH CALLE JIMÉNEZ
tania.calle@epn.edu.ec

CODIRECTORA: PHD. SANDRA PATRICIA SÁNCHEZ GORDÓN
sandra.sanchez@epn.edu.ec

Septiembre 2023

AVAL DEL DIRECTOR

Como directora del trabajo de titulación **DISEÑO DE UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA GESTIÓN DE LA FLOTA VEHICULAR DE UNA EMPRESA PÚBLICA** desarrollado por **CRISTIAN ROBERTO BENALCÁZAR DE LA CRUZ**, estudiante de la Maestría en Sistemas de Información con mención en Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la defensa oral.

Firmado electrónicamente por:
TANIA ELIZABETH CALLE JIMENEZ

PhD. Tania Calle

DIRECTORA

AVAL DEL CODIRECTOR

Como codirectora del trabajo de titulación **DISEÑO DE UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA GESTIÓN DE LA FLOTA VEHICULAR DE UNA EMPRESA PÚBLICA** desarrollado por **CRISTIAN ROBERTO BENALCÁZAR DE LA CRUZ**, estudiante de la Maestría en Sistemas de Información con mención en Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la defensa oral.

Firmado electrónicamente por:
SANDRA PATRICIA
SANCHEZ GORDON

PhD. Sandra Sánchez
CODIRECTORA

DECLARACIÓN DE AUTORÍA

Yo, Cristian Roberto Benalcázar De la Cruz, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según la Ley de Propiedad Intelectual, por su Reglamento y por la Normativa Institucional vigente.

Cristian Roberto Benalcázar De La Cruz

DEDICATORIA

A quienes hicieron posible este logro,

A mi esposa Victoria, por su amor y comprensión, por su aliento y compañía en este camino; y, por recordarme el valor de la perseverancia.
A mi hijo Luciano, por ser mi fuente de inspiración y motivo para esforzarme cada día.

Que este logro nos permita continuar con una vida llena de éxitos y felicidad.

Gracias por ser mi fuerza y mi motivación.

AGRADECIMIENTOS

A la Escuela Politécnica Nacional, en especial a quienes tuve el agrado de tener como mis profesores, por los conocimientos brindados y aportar con su experiencia a ampliar mi visión técnica del mundo de los datos.

A mi directora, PhD. Tania Calle; y, mi codirectora, PhD. Sandra Sánchez por su acertada guía y apoyo constante durante el desarrollo de este estudio.

A la Empresa Pública que me abrió las puertas de su organización para poder realizar el presente trabajo, en particular a las autoridades que visionan una mejora para la Institución con la aplicación de conocimientos y herramientas modernas.

A mi esposa, por escuchar, leer y aportar a este proyecto desde su visión y conocimientos de la organización de la cual se desarrolló el estudio. A mi hijo, por inspirarme a ser mejor cada día.

A mis padres y hermanos, por su apoyo constante a lo largo de los proyectos que me he trazado en la vida, por forjar los cimientos que me han permitido cosechar varios logros; y, por saber estar presentes, cada uno a su manera, pese a la distancia.

A mis suegros, por su constante preocupación y motivación en los proyectos emprendidos; de manera especial a Raquel por el tiempo dedicado para leer el presente trabajo. A mis cuñados por su apoyo, que traspasa incluso el Atlántico.

A mis amigos Xavier y Néstor, por motivarme a su manera a terminar esta fase del camino académico.

A todos quienes aportaron desde su espacio a la consecución de este logro y comparten la alegría de la culminación de esta etapa.

ÍNDICE DEL CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

El presente estudio presenta el diseño de un modelo de aprendizaje automático para gestión de la flota vehicular de una Empresa Pública, centrado en la clasificación de conductores de acuerdo a su nivel de riesgo durante la conducción, para lo cual se analizan características y comportamiento de los conductores y el uso de los vehículos para generar a su vez recomendaciones de optimización de la flota vehicular.

En el desarrollo de este trabajo tiene un papel importante el uso de *web scraping* para la obtención de datos, tanto de los conductores como de la flota vehicular y su uso, los cuales a su vez facilitan el proceso de control de vigencia de la matrícula vehicular realizado tradicionalmente de manera manual por la Empresa.

Es así que, en el Capítulo I se enuncian las generalidades de este proyecto, objetivos que persigue el mismo, su justificación e importancia, se muestra una revisión literaria sobre el *web scraping*, modelos de aprendizaje automático y Sistemas de Información (SI).

En el Capítulo II se describe la metodología a utilizar, se realiza un análisis descriptivo de la flota, características y comportamiento de los conductores, el uso de los vehículos por gerencias y departamentos para generar recomendaciones de optimización de la flota, comprende además la generación de modelos de clasificación de los conductores; y, finalmente se desarrolla un tablero para compartir resultados del análisis de la flota, conductores y modelo generado, además de un mapa para visualizar el uso de los vehículos.

En el Capítulo III se presentan y analizan los resultados obtenidos.

Finalmente, en el Capítulo IV se detallan las conclusiones derivadas del desarrollo del proyecto.

Palabras clave: Aprendizaje automático supervisado y no supervisado, *Web scraping*, Árboles de decisión, Máquinas de soporte vectorial, *K-means*, Agrupamiento jerárquico, Sistemas de Información.

ABSTRACT

This study presents the design of a machine learning model for managing the vehicle fleet of a Public Enterprise, focused on classifying drivers according to their level of risk during driving. To achieve this, driver characteristics, their behavior and the use of vehicles are analized to generate optimization recommendations for the fleet.

Web scraping plays an important role in the development of this work, for obtaining data from drivers, vehicle fleet and its usage, which in turn facilitates the traditionally manual process of vehicle registration validity control performed by the Enterprise.

In Chapter I, the generalities of this project are stated, its objectives, justification, and importance, and a literature review on web scraping, machine learning models and Information Systems (IS) are presented.

In Chapter II, the methodology to be used is described, a descriptive analysis of the fleet, driver characteristics and behavior, and vehicle usage by departments and management to generate fleet optimization recommendations. It also involves the development of driver classification models and finally, a dashboard is developed to share the results of the fleet analysis, driver analysis, generated model and a map for visualizing vehicle usage.

In Chapter III, the results obtained are presented and analyzed.

Finally, in Chapter IV, the conclusions derived from the project development are detailed.

Keywords: Supervised and unsupervised machine learning, Web scraping, Decision trees, Support Vector Machines, K-means, Hierarchical clustering, Information Systems.

1. INTRODUCCIÓN

En la actualidad, pese al vertiginoso avance de la tecnología e incluso la utilización de la inteligencia artificial para automatizar tareas que antaño eran impensables que las realice una máquina, existen aún en las organizaciones, sobre todo de carácter público, actividades que se llevan a cabo de manera manual, lo que conlleva a que la gestión pueda ser catalogada como rudimentaria en ciertas áreas de las instituciones, al no capturar y utilizar adecuadamente todos los datos que pueden generarse. Es así que, una de las áreas que presenta esta problemática es la encargada de la gestión de flotas vehiculares en empresas públicas, existiendo allí una gran oportunidad de mejora.

Con frecuencia los cambios en las flotas de transporte se realizan en base a criterios subjetivos, sin un análisis técnico que los respalde, los cuales van de la mano con débiles estrategias de control y seguimiento de la operatividad y cumplimiento, en donde priman actividades realizadas de forma manual, lo cual da espacio para el error humano [1].

Esto sin duda dificulta la medición del cumplimiento de las actividades de la flota al no contar con información precisa, confiable y oportuna, limitando de manera importante la velocidad de respuesta y una adecuada toma de decisiones, ocasionando con ello un enorme riesgo para las empresas [2].

La estandarización y automatización son clave para una correcta gestión de flotas, en este punto, es imprescindible el uso de un servicio de monitoreo satelital GPS [3] desplegado en plataformas *cloud*. Sin embargo, la ubicación en tiempo real pasó a ser una necesidad secundaria, ya que actualmente no solo se requiere conocer la ubicación de un determinado vehículo y su trayecto óptimo sino gestionar flotas de gran tamaño, siendo de vital importancia la generación de reportes con información clave para el negocio, los cuales a su vez permitan generar alertas tempranas ante cualquier evento que pueda tener un impacto negativo; en dependencia del tipo de flota y negocio incluso se vuelve necesario el uso de las tecnologías IOT[1] para un control mucho más minucioso [4].

En este sentido, la optimización de flotas permite incrementar la eficiencia de todos los procesos empresariales en los que intervienen vehículos, mediante la obtención, integración y análisis de datos en tiempo real del vehículo y su alrededor, incluyendo datos como posición, carga, conductor, velocidad, dirección, estado del vehículo y de la vía, meteorología, tránsito, entre otros [5] [6].

Otra de las principales ventajas de la gestión eficiente de flotas es la reducción de costos, dada principalmente por la detección del uso inadecuado y/o subutilización de vehículos,

[1] El Internet de las cosas (IoT) se compone de dispositivos inteligentes conectados a una red, que envían y reciben datos hacia y desde otros dispositivos. IoT crea nuevas oportunidades para recopilar datos del mundo que nos rodea y administrar una gran cantidad de dispositivos en diferentes ubicaciones físicas. IoT también presenta los desafíos de recopilar, almacenar y analizar enormes volúmenes de datos [28].

pudiendo reducir el número de vehículos y conductores necesarios, sin afectar al cumplimiento de objetivos, así también permite un control en el consumo del combustible y aporta a una planificación óptima evitando el tráfico [7].

Es así que, optimizar las flotas e identificar los principales factores que determinan un mayor nivel de riesgo durante la conducción, es clave para generar estrategias y herramientas para la correcta toma de decisiones en la gestión empresarial.

La Empresa Pública objeto de este estudio, al carecer de un correcto control y análisis del parque automotor, limita de manera importante su accionar y pone en riesgo el cumplimiento de varias de sus funciones, como el proceso de asignación y actualización del parque automotor, y el plan de mantenimiento, generando un impacto negativo para la Gerencia de Administración y Logística, y para toda la organización.

Sobre este punto, la Contraloría General del Estado, en el "Reglamento Sustitutivo para el Control de los Vehículos del Sector Público y de las Entidades de Derecho Privado que disponen de Recursos Públicos" [8], señala:

> *Art. 7.- Registros y estadísticas. - La unidad encargada de la administración de los vehículos, para fines de control y mantenimiento, deberá llevar los siguientes registros:*
> *(...)c) Control de vigencia de la matrícula vehicular, así como, del pago de la tasa por concepto del Sistema Público para Pago de Accidentes de Tránsito.*
> *(…) e) Informes diarios de movilización de cada vehículo, que incluya el kilometraje que marca el odómetro.*
> *(...) i) Registro de entrada y salida de vehículos.*
> *(...) k) Actas de entrega recepción de vehículos.*
> *Art. 11.- Distribución de los vehículos. El encargado o responsable (...) debe asignar las unidades automotrices con criterio técnico y atendiendo las necesidades institucionales.*

Para dar cumplimiento a ello, el Instructivo para la Administración y Control del Parque Automotor de la Empresa, señala lo siguiente:

> *Art. 32.- Responsabilidades de la Unidad Transportes. – (…) para la Unidad de transportes se establecen las siguientes responsabilidades:*
> *(...) c) Emitir informes mensuales o cuando la Gerencia de Administración y Logística considere necesario, sobre el uso de los automotores de la Empresa y arrendados de conformidad con la programación aprobada por la Jefatura del Departamento Servicios Generales.*

Así también, el Reglamento Orgánico Funcional de la Empresa establece funciones específicas al respecto, teniendo para:

- La Gerencia de Administración y Logística
  "(…) d) Gestionar la movilidad integral de la Empresa y el uso eficiente del parque automotor de acuerdo a la normativa legal vigente; (…)"

- El Departamento de Servicios Generales
  "(…) c) Supervisar el cumplimiento de las funciones de la Unidad Transportes referente al parque automotor (vehículos, maquinaria y equipo pesado); así como supervisar la información estadística para la toma oportuna de decisiones; (…)"
- La Unidad de Transportes
  "(…) b) Controlar de acuerdo a la Normativa Interna correspondiente y en el ámbito de su competencia, el parque automotor (vehículos, motocicletas, maquinaria y equipo pesado), de propiedad de la Empresa, incluido los arrendados, así como el mantenimiento oportuno de los mismos; y emitir informes sobre el mal uso para que las autoridades competentes de la Empresa determinen responsabilidades;
  c) Mantener un mecanismo de control para la identificación, ubicación y uso de los vehículos, maquinaria y equipo pesado de la Empresa, de acuerdo con la Normativa Interna correspondiente; (…)
  f) Mantener registros actualizados de la base de datos de conductores de los vehículos asignados a la Gerencia de Administración y Logística; (…)"

Actualmente, la Unidad de Transportes de la Empresa lleva a cabo las mencionadas actividades de manera manual, con información que puede considerarse tardía, imprecisa e incluso incompleta. En este sentido, el presente trabajo utiliza técnicas de aprendizaje automático [9] [10] [11] para diseñar un modelo de aprendizaje que permita clasificar a los conductores de acuerdo a su nivel de riesgo durante la conducción en base a los datos de monitoreo satelital de la flota vehicular, características de los conductores y datos públicos que dan cuenta del comportamiento durante la conducción, como son las infracciones de tránsito registradas en la Agencia Nacional de Tránsito (ANT), lo cual, en conjunto con el análisis de uso de la flota, conductores e infracciones aportaría a una oportuna y correcta toma de decisiones en la gestión de la flota, y en la definición de estrategias para reducción de riesgos durante la conducción y el cumplimiento de normativa.

1.1. Objetivo general

Diseñar un modelo de aprendizaje automático para gestión de la flota vehicular de una Empresa Pública, centrado en la clasificación de conductores de acuerdo a su nivel de riesgo durante la conducción.

1.2. Objetivos específicos

- Generar un proceso de web scraping para la obtención de datos de vehículos de la Empresa Pública, monitoreo satelital e infracciones de tránsito por conductor y vehículo para el desarrollo del proyecto y para facilitar el control de vigencia de la matrícula vehicular.

- Analizar características y comportamiento de los conductores de la Empresa Pública durante la conducción, así como el uso de los vehículos por gerencias y departamento para generar recomendaciones de optimización de la flota.
- Diseñar un modelo de aprendizaje automático que permita clasificar a los conductores en base a su nivel de riesgo durante la conducción e identificar las principales variables que contribuyen al mismo.
- Visualizar los resultados del análisis de la flota, conductores y modelo generado en un tablero, así también generar un mapa para visualizar resultados del análisis del uso de los vehículos.

1.3. Marco teórico

En esta sección se presentan definiciones y conceptos teóricos, se realiza una revisión literaria sobre el *web scraping*, modelos de aprendizaje automático, sistemas de información y sistemas de información geográfica.

En las ciencias exactas, a menudo las teorías se expresan en forma de relaciones matemáticas entre ciertas variables, estas relaciones conforman un modelo matemático que describe una determinada situación; sin embargo, dada la cantidad de algoritmos a revisar, para facilitar la comprensión se presentan estas teorías únicamente de forma gráfica y descriptiva.

1.3.1. Web scraping

Web scraping o raspado web, hace referencia al procedimiento de extracción automática de datos de sitios web mediante software. Usualmente, este software simula la navegación de un humano en la web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.

En los últimos años se ha evidenciado un incremento importante en el uso de *web scraping* o *web crawling*, llegando a recopilar información casi en tiempo real de una variedad de sitios de distintas temáticas tanto en internet como en la red oscura, llegando a ser una tecnología de gran utilidad por su capacidad de generación de información mucho más precisa y detallada que un registro manual [12].

No se limita únicamente a extraer datos estructurados de texto como HTML, semiestructurados como JSON o XML, sino también se extiende a datos no estructurados como imágenes, audios, videos, entre otros, permitiendo además la interacción automática con la web, consultas, descarga de archivos, captura de pantallas, entre otros, volviéndose una herramienta extremadamente útil en situaciones en las que no se cuente con datos y que éstos los veamos publicados en una web.

Con el paso del tiempo se ha ido desarrollando y perfeccionando en varias herramientas y lenguajes de programación, así también se ha ido dando mayor relevancia a los datos dentro de las organizaciones para una correcta toma de decisiones, evidenciando con esto el alto potencial de generación de valor que sostiene.

Es conocido por todos el mayor scraper a nivel mundial, Google, que se encuentra escaneando todo el internet, extrayendo información de cada página web y actualizando permanentemente su índice de búsqueda, el cual a su vez es el sitio web más visitado durante el último año (diciembre de 2021 y noviembre de 2022), dato que da cuenta del valor de los datos y de la gran utilidad del *web scraping* y sus aplicaciones.

Podría decirse que todos hemos sido scrapers, ya que en algún momento hemos realizado un copiar y pegar de alguna web a un documento y lo hemos almacenado, es la idea básica y tradicional del *web scraping*; sin embargo, esto no es escalable, es decir, por ejemplo si estamos trabajando de esta forma con miles de consultas de una web en un período definido y se requiere pasar de miles a millones de consultas en el mismo período sin opción a incrementar personal, se torna una tarea imposible de realizar.

1.3.1.1. Tipos de técnicas y tecnologías

Existen varios tipos de técnicas para realizar *web scraping*, en dependencia de la tecnología empleada y su nivel de automatización, así también la existencia de restricciones desde la web a scrapear, para una simplificación de este tema reducimos los tipos a las siguientes categorías, dejando de lado el tradicional “copiar y pegar – humano” ya que es una tarea manual y no es escalable, aunque en ocasiones es la única opción viable.

- **Agrupamiento de texto y uso de expresiones regulares RegEX:** Esta técnica está basada en el comando UNIX[2] ‘grep’[3]; el cual busca una expresión regular definida por el usuario (por ejemplo, una cadena de texto con características específicas, esto puede ser: de una determinada longitud, que inicia, termina o contiene determinados caracteres) dentro de la fuente y devuelve las líneas coincidentes con dicha expresión; sin embargo, no se recomienda utilizarlas ya que puede generar inconsistencias.
- **Programación del protocolo de transferencia de hipertexto (HTTP):** Esta es la primera técnica que solo es aplicable en la web. Al enviar una solicitud GET a un servidor web, este es capaz de extraer la información de la página web. Las solicitudes HTTP se pueden enviar a través de lenguajes de programación estándar, como Java y Python [13].

[2] Unix es una familia de sistemas operativos caracterizada por ser portable y multitarea.
[3] Búsqueda global de expresiones regulares y líneas coincidentes de impresión

- **Parsers[4] de HTML:** Con esta técnica, la información puede ser extraída haciendo uso del diseño y estructura de la página web. Se dirige principalmente a páginas HTML anidadas [14] y puede utilizarse para varias aplicaciones, como extraer texto, enlaces e información de contacto. Algunos lenguajes, como XQuery y HTQL pueden ser utilizados para parsing de documentos, recuperar y transformar el contenido de documentos HTML.
- **Análisis del modelo de objeto de documento (DOM):** Con esta técnica la estructura de la página web es fácilmente identificable, después de lo cual se puede identificar que elemento contiene la información deseada. XML Path (XPath) es un lenguaje de consulta que se utiliza para extraer información de un DOM.

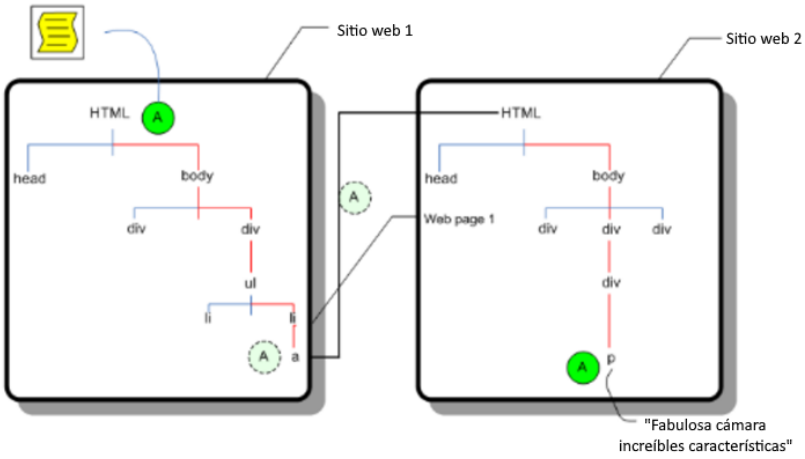


Figura 1 Ejemplo de una consulta de búsqueda XPath *[15]*

En la Figura 1 la línea roja muestra cómo un agente de *web scraping* pasa por un árbol DOM usando XPath. En el documento de la izquierda, la línea termina en un hipervínculo, donde continúa en la página de destino del hipervínculo en el documento correcto abriéndose camino hacia la información deseada [15].

- **Aplicaciones para *web scraping*:** Existen muchas aplicaciones que pueden ser utilizadas para *web scraping*, van desde extensiones web hasta plataformas Saas, estas aplicaciones pueden reconocer automáticamente la estructura de cierta página o brindar una interfaz al usuario donde este pudiera seleccionar los campos que son de interés dentro del documento. De esta forma no es necesario escribir manualmente código para realizar estas tareas.
- **Programar el algoritmo de *web scraping*:** Es una solución a medida, ya que se desarrolla el algoritmo en un lenguaje de programación o framework específico, considerando la estructura e interacción necesaria en la web para la captura de datos,

[4] Parser: Analizador sintáctico, es un programa informático que analiza una cadena de símbolos según las reglas de una gramática formal.

es una solución que aporta mayor flexibilidad y con mayor integración; sin embargo, requiere de amplio conocimiento y experiencia, es la técnica a emplear en el presente estudio en combinación con DOM.

- **Visión por computadora y algoritmos de minería de datos:** Al escanear visualmente páginas web como una persona, la visión artificial y el aprendizaje automático se utilizan para descubrir y almacenar información, esto se realiza en base a grandes colecciones de páginas de similares categorías, para las cuales usualmente existe un *script* o una plantilla, un programa detecta estas plantillas en un contexto específico y extrae su contenido.

1.3.1.2. Herramientas y lenguajes

Para desarrollar un proceso de *web scraping* a medida generalmente se utiliza el lenguaje de programación Xpath, el cual permite construir expresiones que recorren y procesan un documento XML, con una lógica similar a las expresiones regulares para seleccionar partes de un texto sin atributos, el cual permite además buscar y seleccionar teniendo en cuenta la estructura jerárquica del XML.

Ahora, de ser necesaria una interacción con el sitio web para generar la captura de datos, es necesario utilizar el lenguaje OXPath, una extensión del XPath, el cual agrega varias funciones, como acciones del usuario, como clicks y filtrado por características visuales expuestas desde CSS.

Además del lenguaje señalado es necesario contar con conocimientos específicos relacionados y herramientas adicionales para el análisis de los sitios web, la captura y organización de datos como se muestra en la Figura 2.

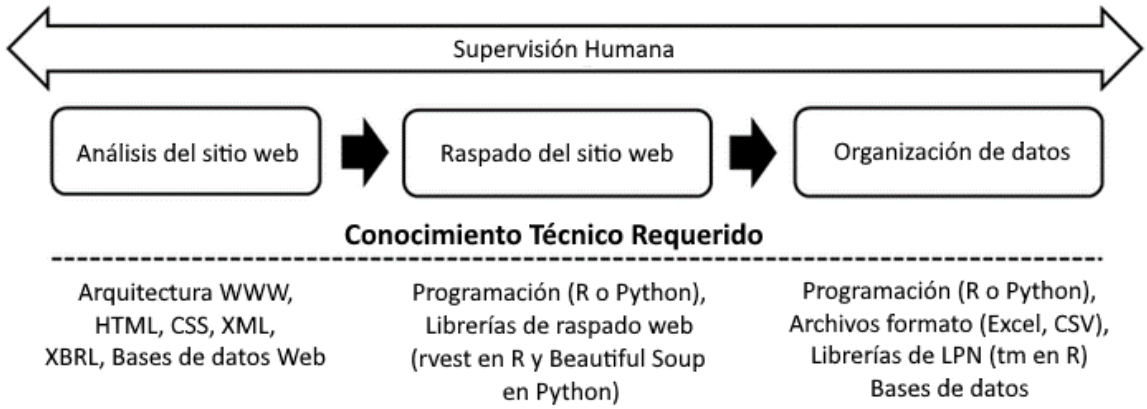


Figura 2 *Web Scraping [16]*

También existen proyectos que utilizan OXPath para una extracción de datos no supervisada con técnicas modernas, enfrentando desafíos como los captchas, limitación de velocidad, entre otros [16].

1.3.2. Modelos de aprendizaje automático

Machine Learning o aprendizaje de máquina o aprendizaje automático es un término que hace referencia al uso de modelos matemáticos con la capacidad de identificar patrones complejos en una gran cantidad de datos y generar predicciones de comportamientos futuros, a partir de éstos.

Los modelos de aprendizaje automático son ampliamente utilizados en empresas que han adoptado o están adoptando la transformación digital, ya que mejora los procesos, proporciona eficiencia productiva, aumenta los ingresos y reduce sus costos.

Se ha observado la aplicación de las técnicas o algoritmos de aprendizaje automático en varias problemáticas del sector empresarial, principalmente en aspectos ligados a la optimización de recursos [17], automatización de procesos [18], segmentación y prospección de clientes [19], entre otros; en general, modelos ligados a estrategias de optimización, captación, retención y rentabilización, la generación de modelos de clasificación de conductores [20] sin duda no es un tema nuevo, pero es importante generarlo en base a las particularidades de cada organización y su necesidad.

1.3.2.1. Tipos de algoritmos

Tradicionalmente se clasificaba a los algoritmos en supervisados y no supervisados, básicamente por la presencia o no de etiquetas o grupos de salida respectivamente; sin embargo, a estos grupos se suman ahora el aprendizaje semisupervisado, el cual es una combinación de los antes señalados.

Además, está el aprendizaje por refuerzo, el cual analiza cómo se comporta el entorno mediante recompensas o castigos, derivados del éxito o del fracaso respectivamente, de esta manera va aprendiendo y buscando la función de valor que maximice la recompensa.

Así también la transducción, similar al aprendizaje supervisado, con la diferencia que éste no genera funciones de manera explícita, sino trata de predecir ejemplos futuros.

Y el aprendizaje multitarea, el cual emplea conocimiento aprendido previamente, donde se es propenso a enfrentarse a diversos problemas.

Además de esta clasificación, los algoritmos de aprendizaje automático se pueden agrupar de acuerdo a sus canales y funciones primordiales, el entrenamiento y la predicción, en este sentido se identifica el aprendizaje automático tradicional y el adaptativo.

- En el aprendizaje tradicional, en la fase de entrenamiento se canalizan los datos recopilados, se depuran, agrupan y transforman; en la segunda función se analiza la

información para generar una predicción que permita tomar decisiones acertadas. Estas funciones si bien están asociadas se tratan de dos entes.

- El aprendizaje adaptativo en cambio emplea un solo canal de entrenamiento y predicción, así los datos se procesan tan pronto como llegan y se generan conclusiones de manera inmediata, y a su vez contempla eventos que pueden alterar el comportamiento en tiempo real para mantener su precisión en todo momento. Siendo mucho más flexible que el tradicional y evitando que el aprendizaje se vuelva obsoleto.

Una vez descritas estas formas generales de clasificación de los algoritmos, es importante señalar una agrupación en función de lo que los algoritmos aprenden, aquí se tienen tres grupos:

- **Clasificación:** Corresponde a técnicas que entrenan modelos que nos permiten predecir la pertenencia a una categoría.
- **Regresión:** Técnicas que permiten predecir un resultado numérico.
- **Aprendizaje de similitud:** Permiten descubrir las formas en que las observaciones en su conjunto de datos se parecen y difieren entre sí.

1.3.2.2. Algoritmos de aprendizaje automático - clasificación

Alineados a los objetivos del presente estudio, nos enfocamos en algoritmos de aprendizaje automático de clasificación, los cuales buscan asignar una categoría a los objetos de análisis, en este caso puntual a los conductores de la flota vehicular de la Empresa.

Los algoritmos de clasificación de aprendizaje automático no supervisado a utilizar en este estudio son agrupamiento jerárquico y *K-means*; y, dentro de los modelos supervisados se utilizan árboles de decisión y máquinas de soporte vectorial.

Para los modelos de aprendizaje supervisado se emplean grupos de entrenamiento y de prueba, generalmente se utiliza un 70% y 30% respectivamente, en dependencia de los datos y modelos se puede modificar estos valores a fin de evitar un sobreajuste.

1.3.2.2.1. Agrupamiento jerárquico

El agrupamiento jerárquico o *Hierarchical Clustering*, es un algoritmo para agrupar datos basándose en la similitud de los mismos, para esto se genera una estructura parecida a un árbol, llamada dendograma, se puede apreciar un ejemplo en la Figura 3, la cual además de mostrar la similitud en base a la distancia, muestra jerarquías en base a la altura de los nodos.

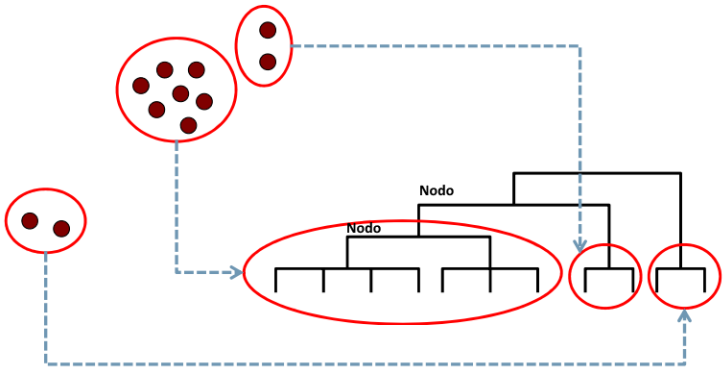


Figura 3 Agrupamiento Jerárquico - Dendograma

No requiere una definición de la cantidad de grupos, éstos se generan de acuerdo a los datos. Para la generación de jerarquías se puede trabajar de manera aglomerativa o divisiva como se muestra en la Figura 4.

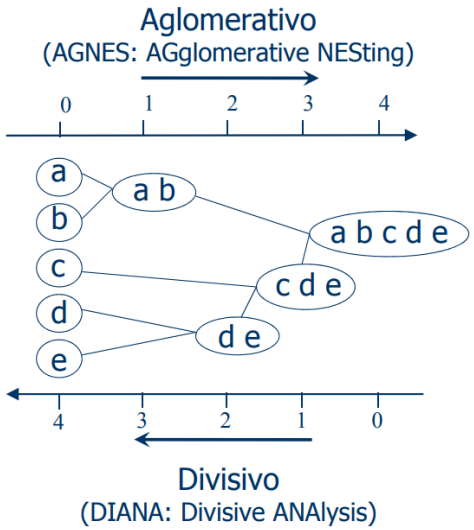


Figura 4 Tipos de jerarquías

La aglomerativa parte de cada punto, lo trata como un grupo y a medida que van relacionándose con otros puntos se van combinando hasta alcanzar el nivel óptimo.
La divisiva en cambio, parte de un solo grupo que contiene a todos los puntos y en cada paso, va partiéndose hasta que cada grupo contenga un punto o se alcance el nivel deseado.
Para la generación de jerarquías además de la dirección se debe considerar el tipo de medida de la distancia entre grupos, pudiendo emplearse varios tipos, entre los más utilizados están:

- Conexión completa: La distancia entre los dos puntos más lejanos de cada grupo.
- Conexión simple: La distancia mínima entre dos puntos de cada grupo.
- Distancia entre medias: La distancia entre las medias de cada grupo.
- La distancia promedio entre pares: Promedio entre todas las distancias que se pueden obtener entre todos los pares de puntos.
- Centroides: La distancia entre los centroides de los grupos.

Ventajas

- Fácil de comprender.
- No es necesaria la definición de cantidad de grupos.

Desventajas

- No es recomendable con una gran cantidad de datos.

1.3.2.2.2. K-means

K-means es uno de los algoritmos de clasificación más utilizados, el cual se basa en la asignación de grupos en base a la similitud de las características de cada punto, para esto se utiliza la menor distancia entre cada punto, previamente es necesario definir el número de grupos deseados, a partir de los cuales se determinan los centroides de los grupos, con los cuales se asigna un grupo a todos los puntos.

La Figura 5 muestra el flujo de un modelo *k-means*, inicialmente los centroides son elegidos de manera aleatoria, a partir de esto se generan las medidas a cada punto y se van realizando asignaciones de grupos, este proceso se repite hasta alcanzar un óptimo, el cual corresponde a una mayor distancia intergrupal y menor distancia intragrupal, visto de esta manera es un problema de optimización. Luego de varias iteraciones finalmente se determinan los centroides y grupos óptimos.

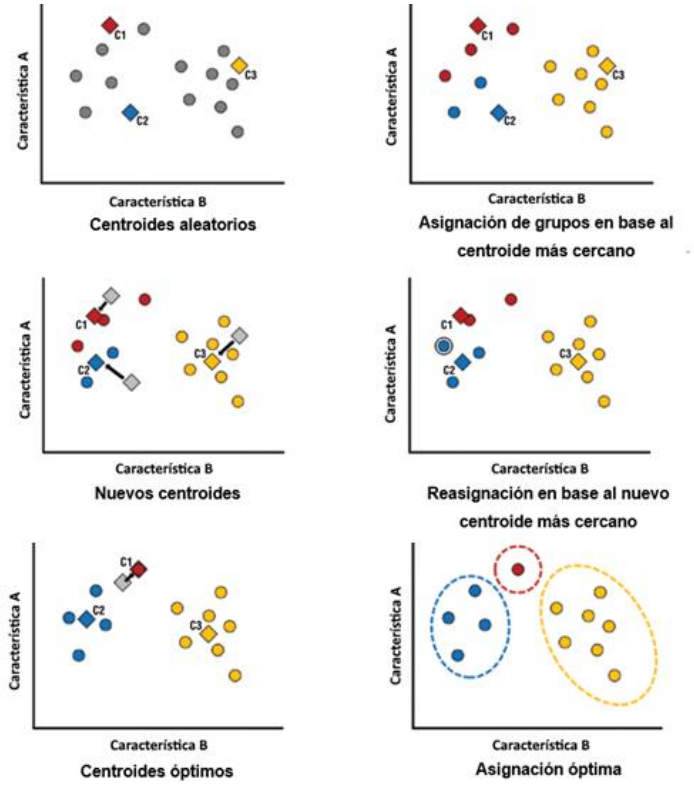


Figura 5 *K-means*

Generalmente se trabaja con la distancia euclídea y el promedio de las distancias, aunque existen derivaciones de este algoritmo para trabajar con la moda o mediana, estos son los llamados K-modas y K-medianas; así también, existen derivaciones como el K-medoids que en lugar de utilizar los centroides utiliza los propios puntos.

Los grupos generados son independientes, pudiendo cada punto pertenecer únicamente a un determinado grupo.

Ventajas

- Fácil de comprender y de rápida ejecución.
- Buenos resultados con pocos o gran cantidad de datos.

Desventajas

- Es necesaria la definición de cantidad de grupos previo a su ejecución.
- Presenta problemas al trabajar con valores atípicos.

1.3.2.2.3. Árboles de decisión

Este algoritmo utiliza una estructura de árbol para representar la relación entre los predictores y el resultado objetivo, se construye en base a una partición recursiva del conjunto de datos original, generando subconjuntos más pequeños hasta que cada subconjunto sea lo más homogéneo posible.

El árbol de decisión está formado por ramas y nodos, éstos últimos pudiendo ser de distintos tipos, como se visualiza en la Figura 6.

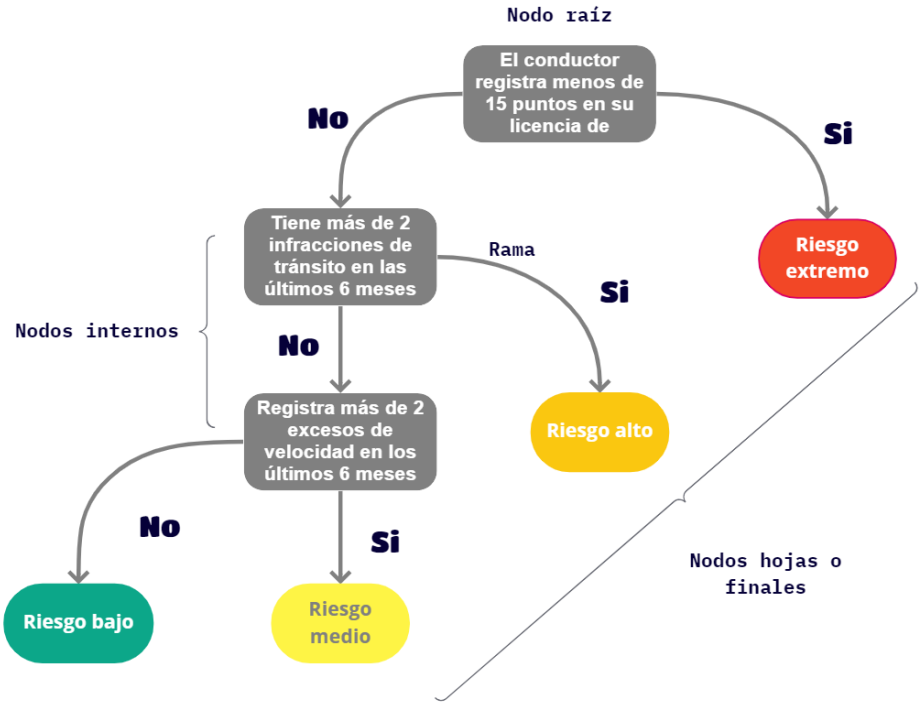


Figura 6 Estructura de un árbol de decisión

- El nodo raíz es el punto de partida, al igual que los nodos internos representan cada una de las características o propiedades a considerar para tomar una decisión.
- Los nodos finales corresponden al resultado de la decisión.
- Las ramas muestran la decisión en función de una determinada condición (por ejemplo: probabilidad de ocurrencia).

La clasificación de todos los datos depende de la complejidad del árbol, en los árboles pequeños es más fácil obtener nodos hoja puros, pero a medida que el árbol crece se vuelve más difícil mantener esta pureza.

Ventajas

- De fácil comprensión.
- Permite identificar la importancia de las variables.
- No se ve influenciado por valores atípicos y valores faltantes (a un cierto grado).
- No existen restricciones por el tipo de datos.

Desventajas

- Puede ocasionar sobreajuste si no se definen límites.
- Es menos preciso que máquinas de soporte vectorial y clasificadores tipo ensamblador (tasas de error 30% más bajas).

1.3.2.2.4. Máquinas de soporte vectorial

Si bien nació como una técnica de clasificación binaria, se ha extendido a problemas de clasificación múltiple, llegando a ser uno de los mejores clasificadores dentro del aprendizaje automático.

Este algoritmo se basa en encontrar un hiperplano que separe de la mejor forma posible las dferentes clases de puntos de datos, esto implica que exista el margen más amplio posible entre las clases, en ejemplo de aplicación de este modelo para una clasificación binaria se puede apreciar en la Figura 7.

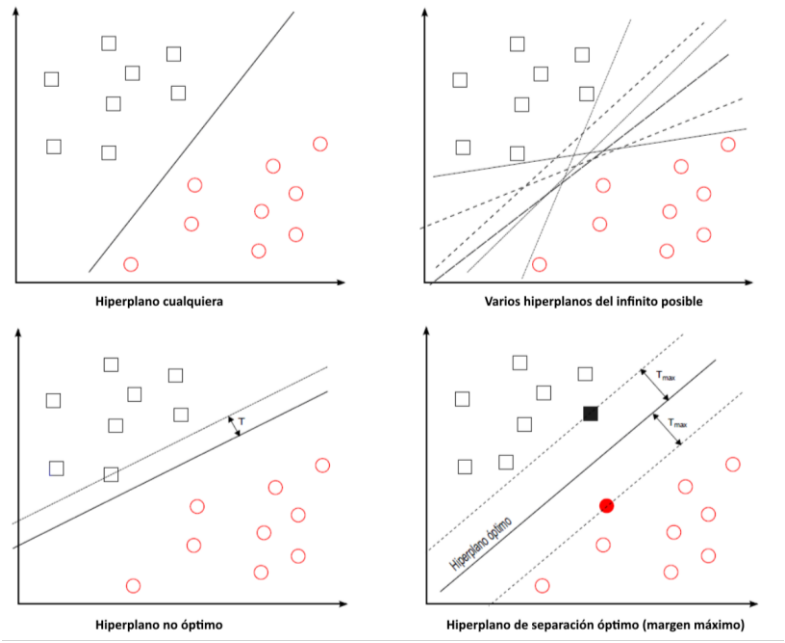


Figura 7 Máquina de soporte vectorial SVM

El algoritmo solo puede encontrar este hiperplano en problemas que permiten separación lineal, si los grupos no son linealmente separables en el espacio original, puede añadirse una tercera dimensión.

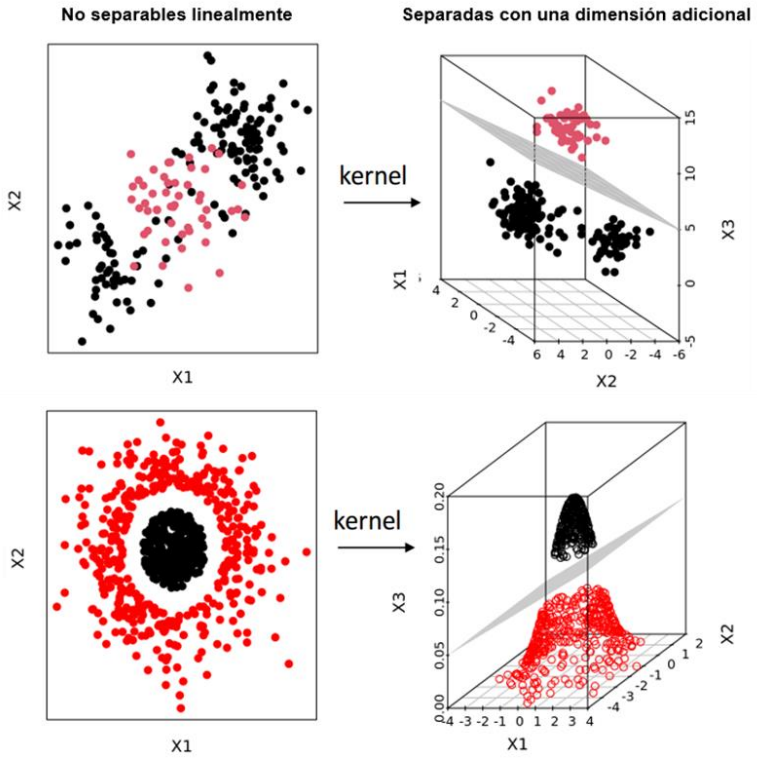


Figura 8 Hiperplanos en 3 dimensiones

La dimensión de un conjunto de datos puede transformarse combinando o modificando cualquiera de sus dimensiones, para transformar un espacio de dos dimensiones en uno de

tres, como el presentado en la Figura 8, donde se utilizan las funciones *kernel*, las cuales devuelven el resultado del producto punto entre dos vectores, generando así un nuevo espacio dimensional distinto al espacio original en el que se encuentran los vectores, gracias a los *kernel*, se puede obtener el resultado para cualquier dimensión.

Existen varios tipos de *kernel*, los más utilizados se presentan en la Figura 9, el *kernel* lineal, polinómico y radial.

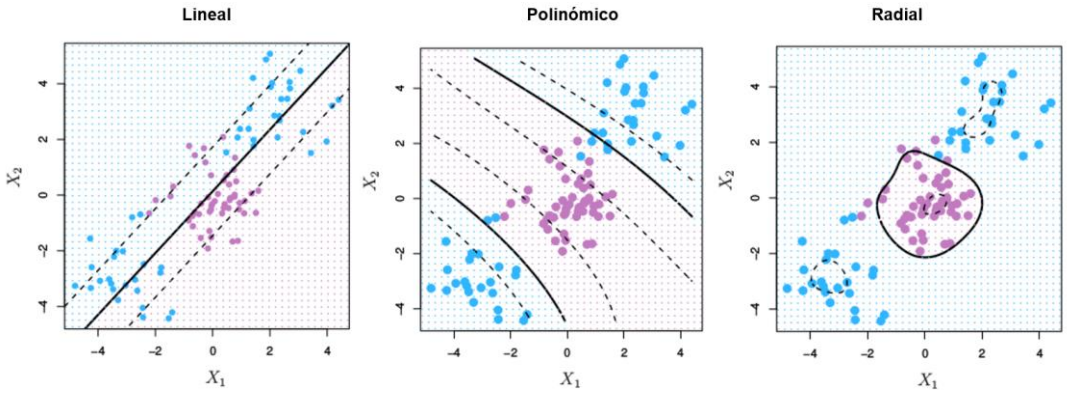


Figura 9 Tipos de *kernel*

Ventajas

- No existen restricciones por el tipo de datos.
- Más preciso que árboles de decisión.
- Menores problemas de sobreajuste ya que en lugar de buscar únicamente el margen de clasificación más ancho posible que consigue que las observaciones estén en el lado correcto del margen, permite que ciertas observaciones estén en el lado incorrecto del margen o incluso del hiperplano.

Desventajas

- Es inestable, un pequeño cambio en los datos puede modificar ampliamente el hiperplano.

1.3.2.3. Evaluación de modelos de clasificación

La evaluación de un modelo de clasificación se da en base a la capacidad que tiene para generalizar el conocimiento implícito en los datos, esto claro está, considerando que ya se hayan solventado posibles problemas de calidad de datos.

La evaluación se realiza para seleccionar el mejor modelo de clasificación para un problema específico, en dicha evaluación se hace una comparativa de las métricas obtenidas de los modelos de clasificación generados utilizando una o varias pruebas estadísticas.

Para este estudio se emplea una matriz de confusión y tasa de aciertos o *accuracy*, a nivel de clases también se observa la sensibilidad y valor predictivo.

Matriz de confusión:

Permite visualizar el número de veces que el modelo ha clasificado de manera correcta o incorrecta.

		Real	
		0	**1**
Predicción	**0**	VN	FN
	1	FP	VP

VN	**Verdadero negativo**
FN	**Falso negativo**
VP	**Verdadero positivo**
FP	**Falso Positivo**

Figura 10 Matriz de confusión

La matriz de confusión presentada en la Figura 10 muestra que siempre es cuadrada debido a que las variables que están en las filas y en las columnas son las mismas. Aquellos elementos que se sitúen en la diagonal son los que el algoritmo ha clasificado de forma correcta. Si se observa una columna, para aquellos elementos que estén en una posición distinta a la diagonal, son el número de veces que el algoritmo ha clasificado una muestra en una clase que no era realmente la suya [18].

Exactitud o *Accuracy*:

Porcentaje de aciertos del modelo, esto es el cociente entre los aciertos y la suma de aciertos y fallos (el total de las predicciones).

$$Exactitud = \frac{\mathrm{VN} + \mathrm{VP}}{\mathrm{VN} + \mathrm{FN} + \mathrm{VP} + \mathrm{FP}}$$

Precisión:

Indica el porcentaje de aciertos de predicciones positivas, es decir, el cociente entre el número de verdaderos positivos y el número de predicciones hechas como positivos (suma de verdaderos y falsos positivos).

$$Precisión = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

Sensibilidad:

Porcentaje de verdaderos positivos. Es decir, es el cociente entre los verdaderos positivos y la suma de todos los positivos que hay.

$$Sensibilidad = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

F1 score:

Es un porcentaje entre medias de la sensibilidad y la precisión, por lo que tiene en cuenta ambos valores. Es por ello que a medida que aumente el valor de F1 es mejor el modelo.

$$F1score = \frac{2 * \text{precisión} * \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}}$$

Es importante señalar que al aplicar una gran cantidad de medidas de rendimiento en la evaluación de modelos el grado de complejidad crece, en especial en los modelos multiclase, por tanto, si bien se generan estas medidas, tienen una mayor relevancia la exactitud o *accuracy* [21].

1.3.3. Sistemas de información SI

En el entorno cada vez más competitivo en el que se desenvuelven las empresas, es primordial generar una rápida respuesta frente a las estrategias de la competencia, para lo cual es necesario contar con información precisa, confiable y oportuna, ante esta necesidad se han desarrollado varias herramientas enfocadas a la gestión de datos y generación de información para la toma de decisiones, ahora conocidas como herramientas de Inteligencia de Negocios.

Con estas herramientas, el seguimiento de las métricas clave para un departamento, gerencia u organización puede plasmarse en un solo panel o tablero y ser accesible en todo momento y lugar, permitiendo evaluar rápidamente el rendimiento de un determinado proyecto, proceso o negocio.

Si bien los sistemas de información nacen como herramientas gerenciales, su aplicación se ha extendido a todos los niveles, pudiendo clasificarlos en tres grupos:

- Operativos: Cuyo objetivo es el seguimiento de los procesos o unidades de negocio de la organización para la toma de decisiones ligadas a su actividad.
- Directivo: Presenta resultados de la organización por sus diferentes áreas, creando una visión general.
- Estratégico: Asociado con la metodología *Balanced Scorecard*, agrupa la información por objetivos, iniciativas e indicadores, para la alta dirección de la organización permitiendo conocer el comportamiento de la estrategia y su ejecución.

Un SI extrae e integra datos de distintas fuentes, realiza transformaciones, genera y consolida información clave para una determinada temática y la presenta en un panel interactivo que brinda al usuario la posibilidad de realizar análisis y comparativas de manera rápida y sencilla. A enero de 2022, la consultora Gartner en su Cuadrante Mágico para Plataformas de Analítica e Inteligencia de Negocios presentado en la Figura 11, sitúa a Microsoft Power BI en primer lugar por cuarto año consecutivo, considerando su integridad de la visión y capacidad de ejecución, reconocimiento logrado también por la amplia comunidad de usuarios, además de las características y funciones del software.

Figura 11 Cuadrante mágico para Plataformas de Analítica e Inteligencia de Negocios

2. MÉTODO

Esta sección comprende la metodología a usar en la extracción de información, generación de los modelos de clasificación de los conductores y en el desarrollo del sistema de información donde se presentan los resultados.

2.1. Web scraping

Para el desarrollo de los procesos de *web scraping* requeridos en este estudio se programan los algoritmos utilizando la técnica de análisis DOM y se utilizan, las siguientes herramientas:

- Rstudio build 576 y R 4.1.1 (dplyr 1.0.9, reticulate 1.2.5, Rselenium 1.7.9, XML 3.99-0.10, xlsx 0.6.5, tidyverse 1.3.2, rvest 1.0.2 y httr 1.4.3)
- Python 3.11.0 (selenium 4.5.0, pandas 1.5.1, datetime, time)
- Firefox Browser 106.0.5
- Chrome Driver Versión 105.0.5195.54

2.1.1. *Web scraping* flota vehicular

Los datos de la flota vehicular de la Empresa presentan 442 registros de vehículos, con los campos detallados en la Tabla 1:

Tabla 1 Campos de tabla de datos de vehículos

Campo	Valores perdidos
Marca	27
Modelo	110
Placa	0
Gerencia	2
Departamento	2
Unidad	3
Clase de vehículo	0
Color	170
Número de chasis	28
Número serie motor	0
Año de construcción	0
Gps	0

Campo	Valores perdidos
Custodio	3
Sitio Parqueo	28
Estado del vehículo	394
Operatividad	395
Propiedad	0
Pool de transportes	0
Baja	0
Gps	0
Numero sap	34
Número activo	38
Valor adquisición	34

De los cuales 47 registros corresponden a maquinaria y vehículos pesados (tractores, cargadoras, rodillos, excavadoras, montacargas, grúas, entre otros), los cuales por sus características no tienen un registro de matrícula; y, dado que su uso se da en sitios y condiciones específicos quedan fuera del presente análisis.

Considerando que la Empresa Pública realiza el proceso de control de vigencia de la matrícula vehicular de manera manual, se genera un proceso de *web scraping* de la página web del Servicio de Rentas Internas (SRI), en la opción de Vehículos, en la consulta de valores a pagar por placa o chasis[5] para enriquecer el análisis y facilitar el control señalado.

El control de vigencia de la matrícula básicamente corresponde a presupuestar el monto a cancelar por motivo de matriculación de la flota vehicular y generar un respaldo de dicho valor por unidad, el cual para tener validez debe señalar de manera expresa su fuente, es así que se necesita una captura de cada consulta, además del valor en una hoja de cálculo.

Para el *web scraping* de la página de consulta de valores a pagar por placa del SRI se desarrolla una función en R “matricula”, la cual tiene como argumentos la placa y un número, para consultar y para ordenar los resultados y capturas respectivamente.

[5] https://srienlinea.sri.gob.ec/sri-en-linea/SriVehiculosWeb/ConsultaValoresPagarVehiculo/Consultas/consultaRubros

Dicha función ingresa a la página web oficial de la institución en mención, como se muestra en la Figura 12, en donde identifica los elementos de la misma (campo de ingreso de placa y botón de consulta), con los cuales interactúa ingresando una placa y consultando.

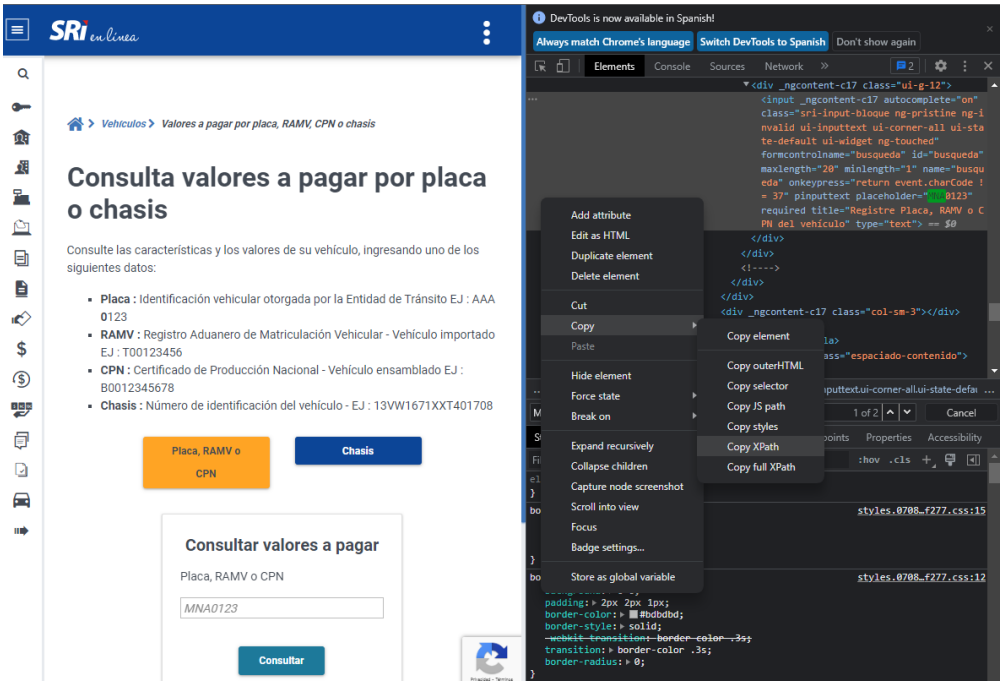


Figura 12 Web SRI a scrapear

Seguido de esto nuevamente se identifican los elementos de la página de resultados (botones de mostrar detalles vehículo y mostrar valores), como se visualiza en la Figura 13, ya que se debe interactuar también con dichos botones para mostrar todos los datos.

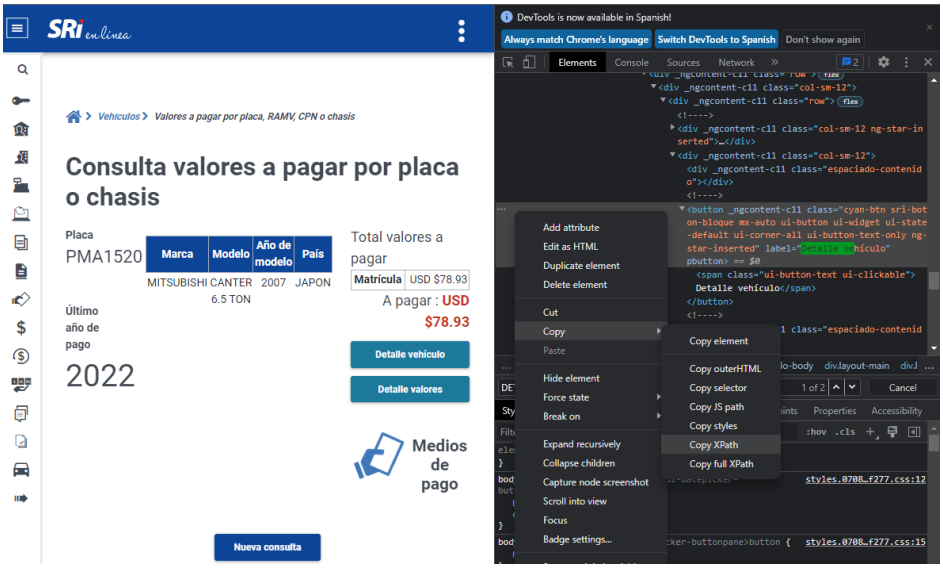


Figura 13 Web SRI – identificación de elementos

Luego de ver todos los resultados en pantalla se realiza una captura de imagen, la cual se almacena y el archivo se nombra con el número y placa, esto se presenta en las Figuras 14, 15 y 16.

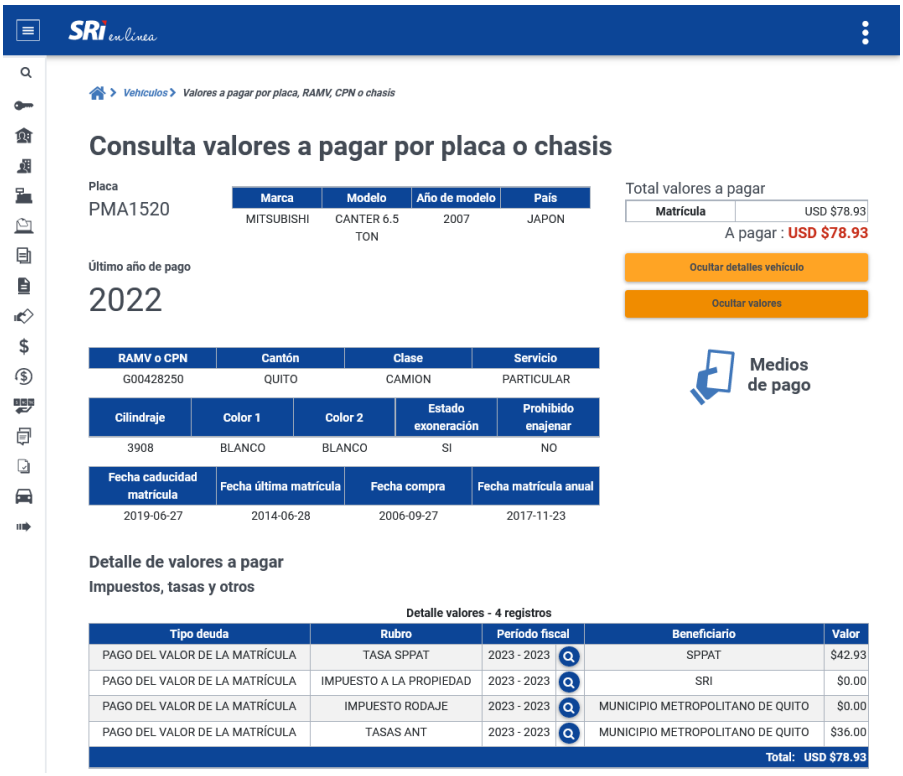


Figura 14 Web SRI – datos a extraer

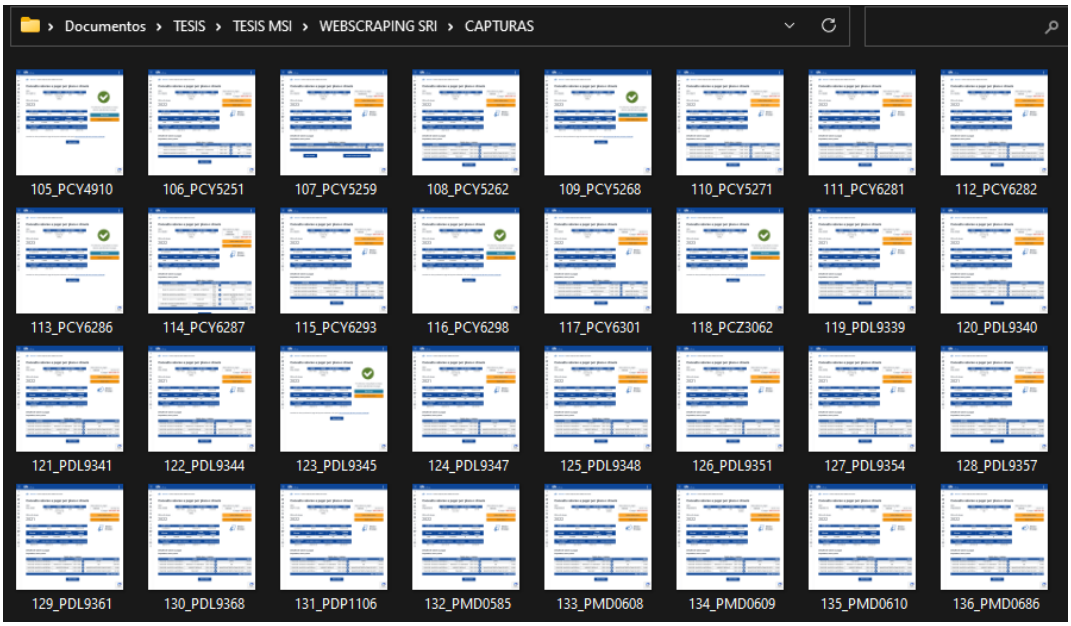


Figura 15 Web SRI – capturas 1

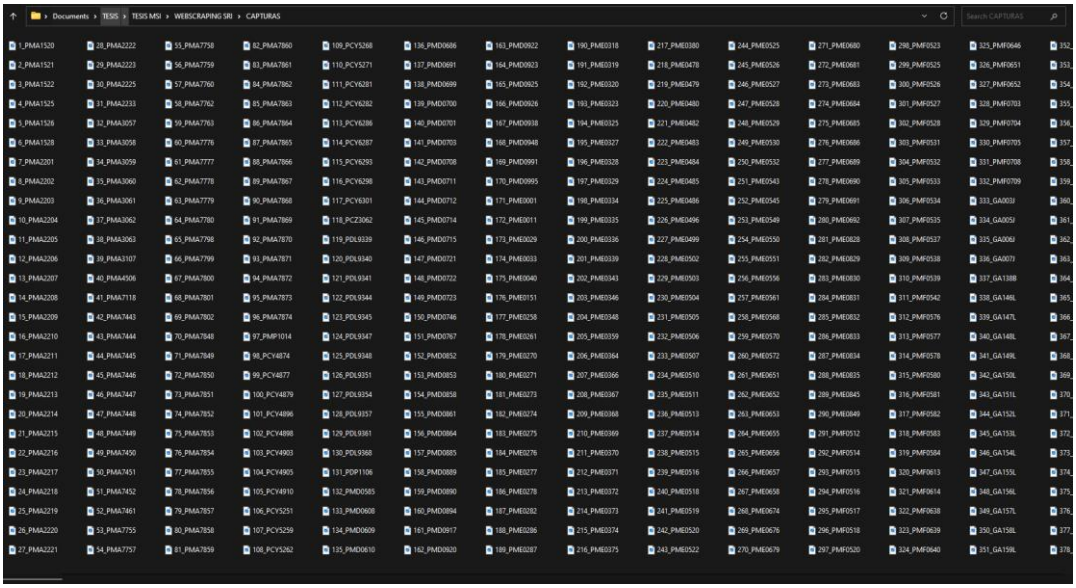


Figura 16 Web SRI – capturas 2

A continuación, se identifican y almacenan todos los datos observados con varias sentencias *for* y matrices creadas para el efecto, con varias reglas de validación.

La función antes descrita se utiliza con una sentencia *for* para todas las placas, al terminar dicha sentencia se tienen las capturas y los datos en dos tablas “datos_flota” y “detalle_pagar” que se exportan en formato “.xlsx” para los análisis posteriores, los campos obtenidos se enlistan en la Tabla 2.

Tabla 2 Campos de las tablas resultantes del *web scraping* de la flota

datos_flota
Placa
Marca
Modelo
Año de modelo
Pais
RAMV o CPN
Canton
Clase
Servicio
Cilindraje
Color1
Color2
Estado exoneracion
Fecha caducidad matricula
Fecha ultima matricula
Fecha compra
Fecha matricula anual
Total a pagar

detalle_pagar
Tipo deuda
Rubro
Período fiscal
Beneficiario
Valor
Placa

Con estos datos se obtiene el valor total a cancelar por concepto de matrícula de toda la flota del año 2023 ($36,194.62), el cual se desagrega por clase de vehículo en la Tabla 3 y por tipo de rubro en la Tabla 4. Para el pago del mismo es necesario contar con la captura del SRI con el valor de cada unidad como requisito y para respaldo de la Empresa.

Tabla 3 Valor a cancelar por concepto de matrícula por clase de vehículo

Clase vehículo	Número de vehículos	Valor matrícula
Camioneta	202	$ 22,767.92
Camion	60	$ 5,447.12
Jeep	39	$ 3,458.67
Tanquero	10	$ 1,186.10
Especial	9	$ 1,067.49
Volqueta	10	$ 894.77
Motocicleta	62	$ 723.86
Omnibus	1	$ 585.95
Automovil	1	$ 62.74
Na	1	$ -
Total	**395**	**$ 36,194.62**

Tabla 4 Valor a cancelar por concepto de matrícula por rubro

Rubro	Valor matrícula
Pago de ajustes	$ 71.79
Pago del valor de la matrícula	$ 35,747.32
Pago del valor de transferencia de dominio	$ 375.51
Total	**$ 36,194.62**

2.1.2. *Web scraping* monitoreo satelital

La Empresa cuenta con el servicio de monitoreo satelital de Omnilogik[6], el cual reporta datos cada minuto cuando las unidades se encuentran en movimiento y cada hora cuando están detenidas. A través de este servicio se puede obtener: la ubicación geográfica, fecha, hora, velocidad y dirección de cada unidad.

Si bien el servicio cumple con la principal función de monitoreo en tiempo real, tiene varias limitantes:

- No muestra indicadores de toda la flota.
- No permite una descarga masiva de los datos.

[6] http://69.64.40.175:8881/

- Las consultas son bastante tardías y al ampliar el tiempo de consulta la página colapsa.
- No tiene una interfaz agradable para el usuario.

Las cuales dificultan de manera importante el análisis de la flota e impiden tomar acciones oportunas.

Para el *web scraping* de la página de monitoreo satelital se desarrolla una función en Python “ruta_fecha_rango”, la cual tiene como argumentos el código SAP, fecha de inicio y fecha final de consulta, para consultar y delimitar el rango de fechas respectivamente.

Dicha función ingresa a la página web de la empresa proveedora, como se muestra en la Figura 17, en donde identifican los elementos de la misma (campo de usuario, clave y botón de ingreso), con los cuales interactúa para ingresar al portal.

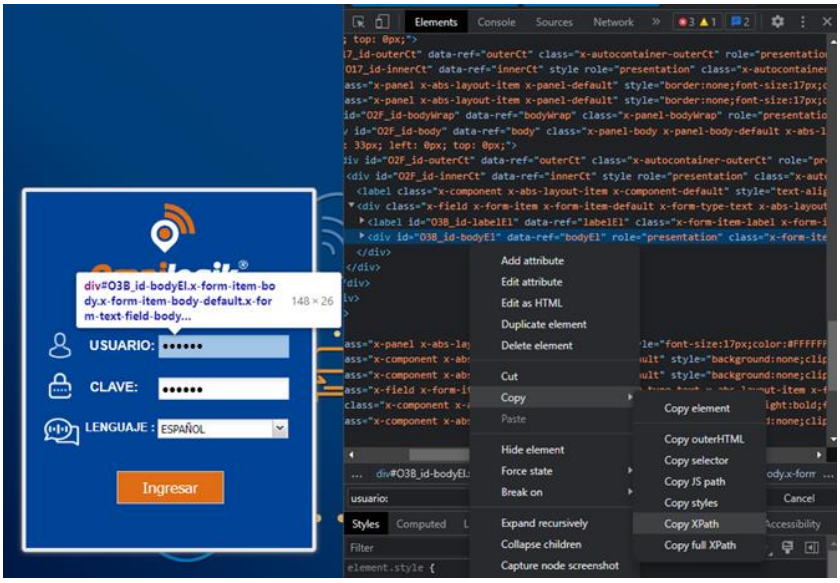


Figura 17 Web Rastreo Satelital a scrapear

Seguido del ingreso al portal, se visualiza la Figura 18, donde se identifican los elementos de la página (campo de búsqueda y botón asociado), ya que se debe interactuar también con dichos botones para mostrar todos los datos requeridos.

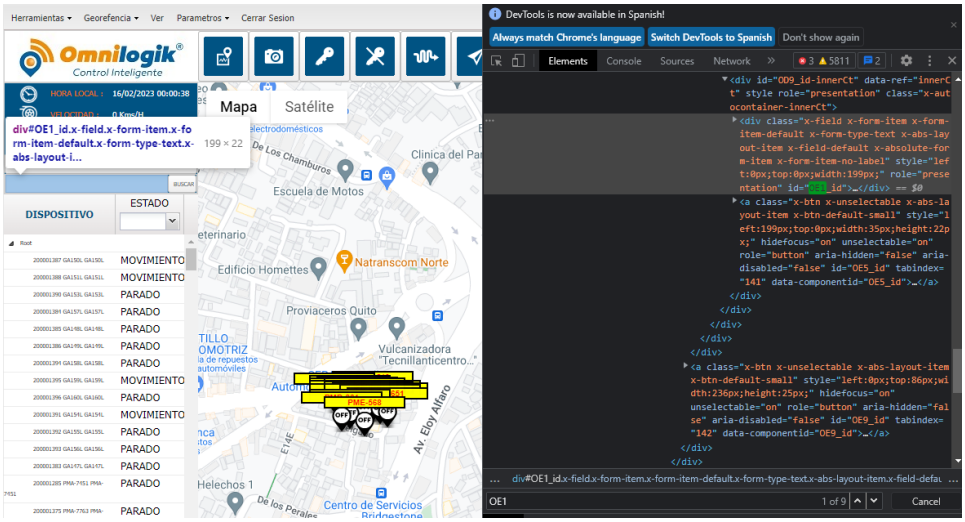


Figura 18 Web Rastreo Satelital – identificación de elementos

Posterior al ingreso y búsqueda de un código SAP, se identifica e interactúa con el botón de reportes, luego dentro del cuadro desplegado se interactúa con los campos de fechas y horas; y, el botón consultar, esto se aprecia en la Figura 19.

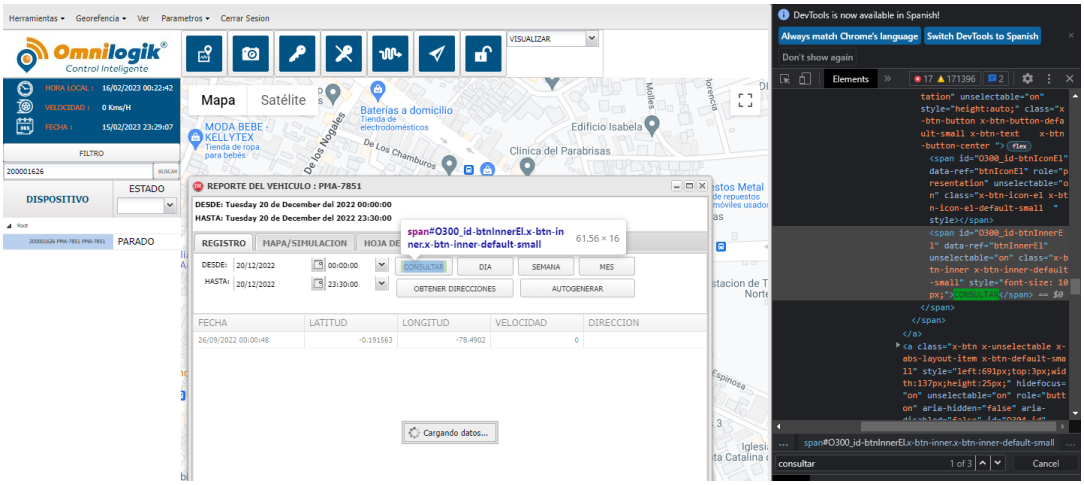


Figura 19 Web Rastreo Satelital – interacción

Seguido de esto se interactúa con el botón “Exportar” y un par de botones adicionales de descarga, en formato CSV y link de descarga, que van apareciendo con las interacciones, como se visualiza en la Figura 20.

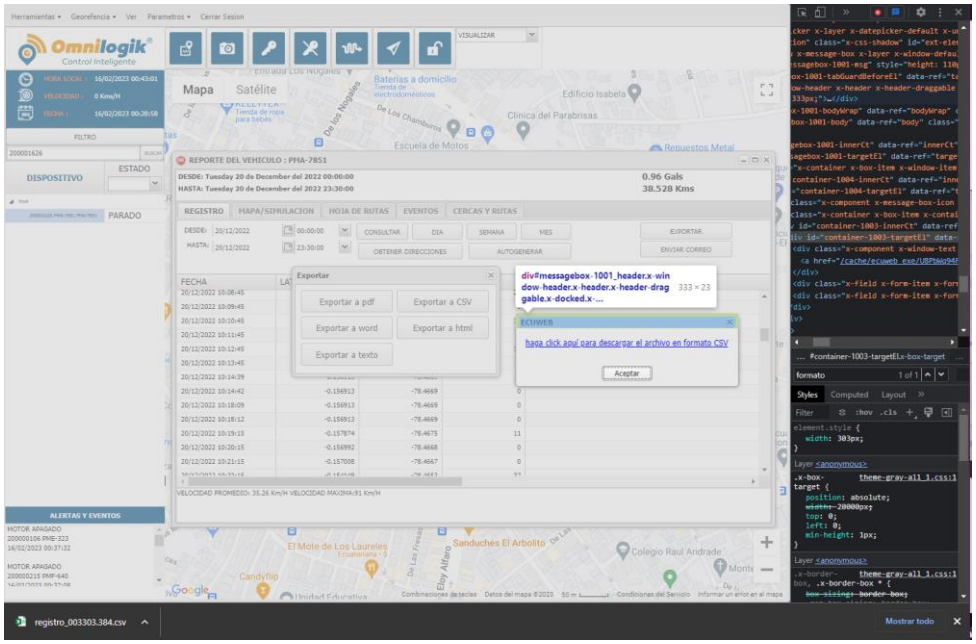


Figura 20 Web Rastreo Satelital – identificación de elementos de descarga

La función descrita se coloca dentro de una sentencia for para la consulta y descarga de los datos del año 2022 para toda la flota vehicular, si bien la función se desarrolla en Python, se la utiliza desde R mediante un *script* adicional, se consideran tiempos de espera para las interacciones y reglas condicionales para evitar errores.

Los resultados de este proceso se consolidan en una tabla con más de dos millones de registros, los campos de la misma se detallan en la Tabla 5:

Tabla 5 Campos de las tablas resultantes del *web scraping* del monitoreo satelital

Campos
Fecha
Latitud
Longitud
Velocidad
Direccion
Id_dispositivo

Campos adicionales
Fecha_ymd
Check_mov
Dis_mtr
Tmp_mins

A partir de esto se añade una marca de movimiento, distancia recorrida en metros y tiempo de viaje en minutos, eliminando además los registros intermedios cuando no existe movimiento de los vehículos para reducir el tamaño de la tabla a 1’437.500 registros.

2.1.3. *Web scraping* infracciones de tránsito vehículos y conductores

Para obtener información de las infracciones de tránsito de la flota vehicular de la Empresa se desarrolla una función de *web scraping* en R, la cual trabaja en la página web de consulta de citaciones[7] de la ANT.

Dicha función identifica los campos de selección de tipo de consulta, ingreso de placa o cédula y botón de consulta, como se visualiza en la Figura 21.

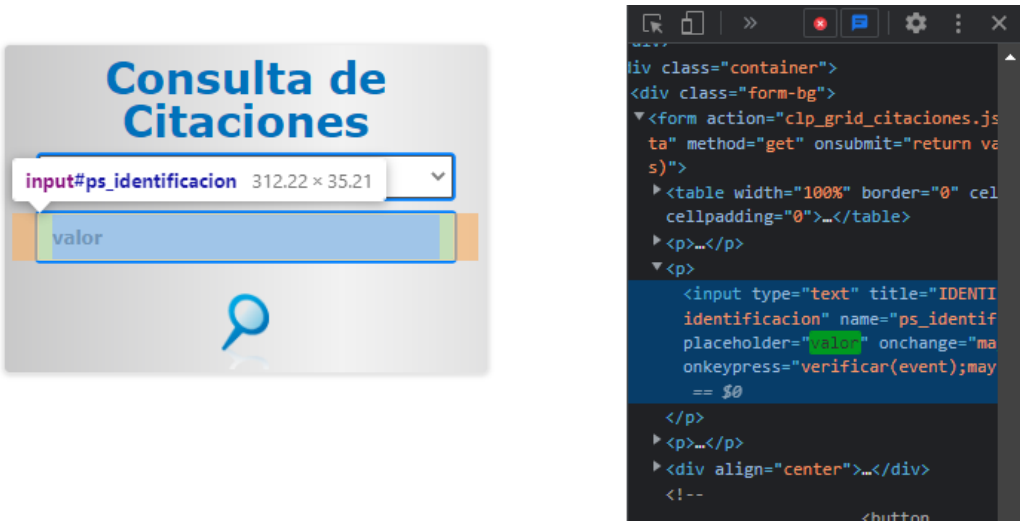


Figura 21 Web Infracciones ANT – identificación de elementos - 1

Posterior a la consulta se identifica el número de filas y páginas en la tabla, así también los botones de estado de las infracciones para ir capturando los datos de cada estado y página, como se presenta en la Figura 22.

Figura 22 Web Infracciones ANT – identificación de elementos - 2

Se utiliza dicha función en una sentencia *for* para consultar las placas de toda la flota, obteniendo una tabla de 545 registros con los campos enlistados en la Tabla 6:

[7] https://consultaweb.ant.gob.ec/PortalWEB/paginas/clientes/clp_criterio_consulta.jsp

Tabla 6 Campos de las tablas resultantes del *web scraping* de las infracciones

Campo
Infraccion
Entidad
Citacion
Placa
Fecha de Emision
Hora Emision
Fecha Notificacion
Hora Notificacion
Puntos
Sancion
Multa
Remision
Total a pagar
Articula/Literal
Bq
Estado

Para la obtención de las infracciones de los conductores de la flota se utiliza la misma función con ligeras variaciones a fin de capturar además el puntaje en sus licencias, los campos extraídos son los mismos de la tabla anterior. Se genera una tabla de 2.621 registros con las infracciones de los conductores desde 1999 al 5 marzo de 2023.

2.2. MODELOS DE CLASIFICACIÓN

Para el desarrollo del presente estudio se utiliza la metodología *Cross Industry Standard Process for Data Mining* (CRISP-DM), la cual es la guía de referencia en el desarrollo de proyectos de minería de datos más utilizada en el mundo.

2.2.1. Metodología CRISP-DM

CRISP-DM proporciona una visión general del ciclo de vida de un proyecto de minería de datos y organiza las tareas de minería de datos en las fases presentadas en la Figura 23.

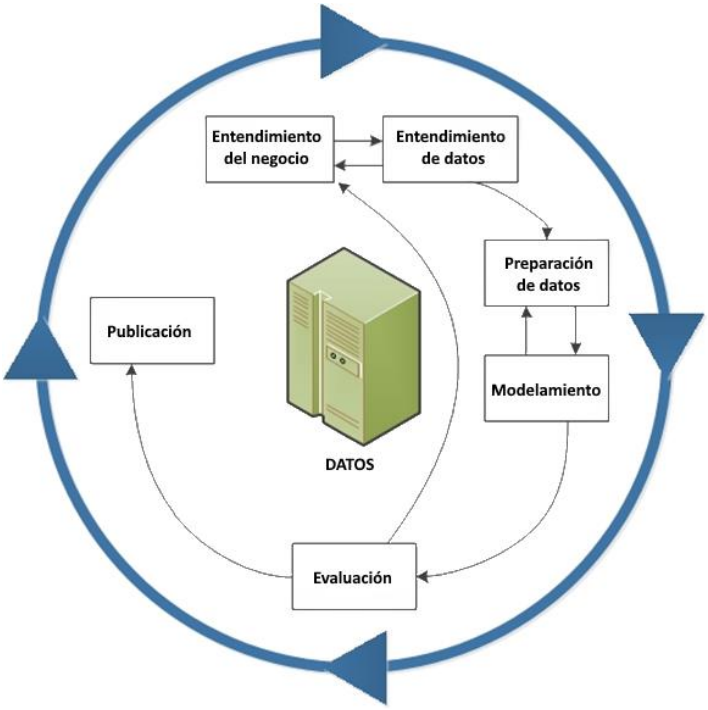


Figura 23 Fases del ciclo de vida de un proyecto de minería de datos CRISP-DM1.0 *[22]*

1. Entendimiento del negocio: Se identifican los antecedentes de la problemática, situación actual, objetivos estratégicos y criterios de éxito.
2. Entendimiento de los datos: Se generan procesos de extracción, descripción, exploración y verificación de calidad de los datos.
3. Preparación de los datos: Se generan procesos de selección, limpieza, transformación e integración de los datos a utilizar.
4. Modelamiento: Se seleccionan las técnicas a utilizar y se generan varios modelos de aprendizaje automático, considerando grupos de entrenamiento y de prueba.
5. Evaluación: Se evalúan los resultados de los modelos en base a distintos criterios de desempeño [23]; así también, se identifican las variables con mayor contribución al modelo, se realiza una revisión del proceso y se determina los próximos pasos.
6. Despliegue (puesta en producción): Se comparten los resultados para el desarrollo de una estrategia [24] basada en datos.

CRISP-DM destaca entre sus ventajas: la facilidad en el entendimiento por parte del cliente de la planificación y ejecución del proyecto, además de ofrecer la posibilidad de replicación de proyectos, con independencia de la industria y herramientas; y, su enfoque en el negocio y en el análisis.

2.2.1.1. Entendimiento del negocio

La gestión de flotas tradicionalmente era un proceso estático, donde las empresas ignoraban en muchas ocasiones la ubicación de sus vehículos, el estado de los mismos y de las mercancías que transportaban, no sólo no tenían estos datos cuando los vehículos estaban en movimiento, sino también cuando estaban estacionados. La fiabilidad de las operaciones en muchos casos se basaba en la memoria de los operadores de las flotas o de los conductores e inventarios, que registraban todos estos datos de manera manual [25].

Con el desarrollo tecnológico observado en los últimos años, pasamos de un control manual a uno digital; sin embargo, muchas veces éste no se adapta a las necesidades específicas de un determinado negocio, generando tareas adicionales, por cuanto, para una gestión óptima de la flota, se debe considerar los requerimientos específicos del negocio y todos los elementos intervinientes, para poder rápidamente dar respuesta a preguntas clave como en el caso puntual de este estudio, identificar a los conductores más seguros, subutilización de los vehículos, la correcta asignación de vehículos y conductores, y presupuestar gastos de matriculación, pudiendo alcanzar también la identificación de rutas y horas óptimas de viaje, la identificación de los vehículos más eficientes, identificación y disminución de tiempos muertos, entre otros.

Como se indicó en secciones anteriores, la identificación de conductores con mayor nivel de riesgo es clave para una correcta toma de decisiones, ya sean de reasignación de vehículos, de capacitación o cambios en el personal.
Un bajo puntaje en la licencia de conducción, así como una alta cantidad de infracciones de tránsito tanto con vehículos de la Empresa como particulares son claros indicios de riesgo durante la conducción, por cuanto deben ser considerados para la clasificación de conductores.

2.2.1.2. Entendimiento y preparación de datos

Esta etapa engloba la comprensión de los datos a utilizar para el modelo, para lo cual se examinaron las bases de datos de la flota de la Empresa, conductores, monitoreo satelital e infracciones de tránsito; y, se realizó un análisis exploratorio, el cual permitió tener un claro panorama de la temática en cuestión. Dicho análisis se presenta en el tablero desarrollado en la sección 3.1.

En relación a la preparación de datos, previamente se generaron procesos de limpieza y transformación, los cuales básicamente corresponden a controles de calidad de datos en identificadores principales como son placas de los vehículos, cédulas de conductores, codificación para controlar caracteres especiales, fechas y horas.

Se generan variables adicionales que resumen la información del comportamiento de los conductores en vehículos de la flota y en vehículos particulares, estas variables se describen en la Tabla 7.

Tabla 7 Variables para generación de modelos

Variable	Descripción
clase_vehiculo	Clase de vehiculo asignado
rango_edad	Rango de edad
nivel_estudios	Nivel de estudios
estado_civi	Estado civil
rango_puntos	Rango de puntos de licencia
flota_2022_N_multas	Número de multas último año (flota)
flota_2022__Puntos	Puntos perdidos en multas último año (flota)
conduc_2022_N_multas	Número de multas último año (particular)
conduc_2022__Puntos	Puntos perdidos en multas último año (particular)
N_multas_cond_sin_lic	Número de multas (flota) por conducir sin licencia
N_multas_desob_orden	Número de multas (flota) por desobedecer agente y/o señal
N_multas_est_sit_proh	Número de multas (flota) por estacionar en sitios prohibidos
N_multas_exc_vel_mode	Número de multas (flota) por exceso de velocidad - rango moderado
N_multas_lic_caduc	Número de multas (flota) por conducir con licencia caducada/anualada/revocada/suspendida
N_multas_cintu	Número de multas (flota) por no utilizar cinturón de seguridad
Puntos_cond_sin_lic	Puntos multas (flota) por conducir sin licencia
Puntos_desob_orden	Puntos multas (flota) por desobedecer agente y/o señal
Puntos_est_sit_proh	Puntos de multas (flota) por estacionar en sitios prohibidos
Puntos_exc_vel_mode	Puntos de multas (flota) por exceso de velocidad - rango moderado
Puntos_lic_caduc	Puntos de multas (flota) por conducir con licencia caducada/anualada/revocada/suspendida
Puntos_cintu	Puntos de multas (flota) por no utilizar cinturón de seguridad
condArt._139_N_multas	Número de multas leves (particular) Art.139
condArt._140_N_multas	Número de multas leves (particular) Art.140
condArt._141_N_multas	Número de multas leves (particular) Art.141
condArt._142_N_multas	Número de multas graves (particular) Art.142
condArt._144_N_multas	Número de multas graves (particular) Art.144
condArt._145_N_multas	Número de multas muy graves (particular) Art.145
condTotal_N_multas	Número de multas (particular)
condTotal__Puntos	Puntos de multas (particular)
Máx._de_Velocidad_maxima_mov	Velocidad máxima
Máx._de_Distancia_mov	Máxima distancia diaria recorrida
Máx._de_Tiempo_h_mov	Máximo tiempo de uso diario

2.2.1.3. Modelamiento

Esta etapa engloba la generación de los modelos de aprendizaje automático, se inició con los modelos no supervisados, agrupamiento jerárquico y *K-means*; y, a partir de esta primera

clasificación se generan modelos supervisados, árbol de decisiones y máquina de soporte vectorial los cuales permiten además identificar las principales determinantes de un mayor nivel de riesgo durante la conducción.

Para el modelamiento se utiliza la librería caret del lenguaje de programación R tanto para los modelos supervisados como no supervisados, esto a su vez facilita la comparativa y evaluación de modelos.

2.2.1.3.1. Modelos no supervisados

Se generan variables *dummy* para las variables categóricas a fin de utilizarlas dentro de los dos modelos.

Agrupamiento jerárquico

La identificación de la jerarquía se realiza de manera divisiva, utilizamos diferentes métodos de conexión, obteniéndose un mejor resultado en el método de conexión completa, en base al cual generamos el dendograma mostrado en la Figura 24, en donde identificamos 6 grupos.

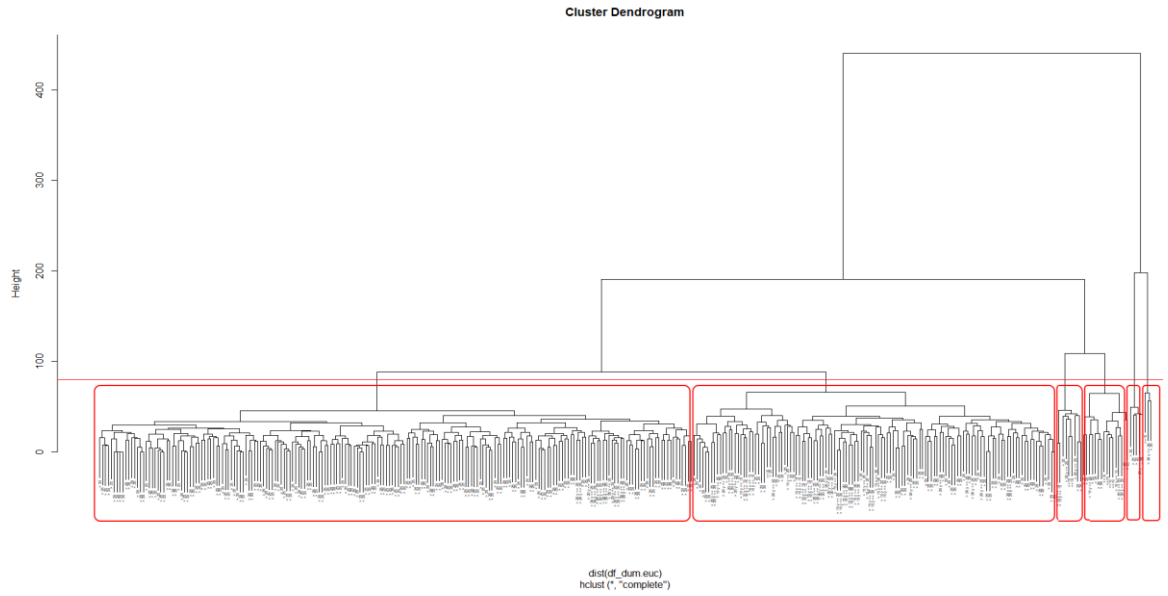


Figura 24 Dendograma de conductores

Se realiza un modelo adicional a partir de los tres primeros componentes principales, este se presenta en la Figura 25.

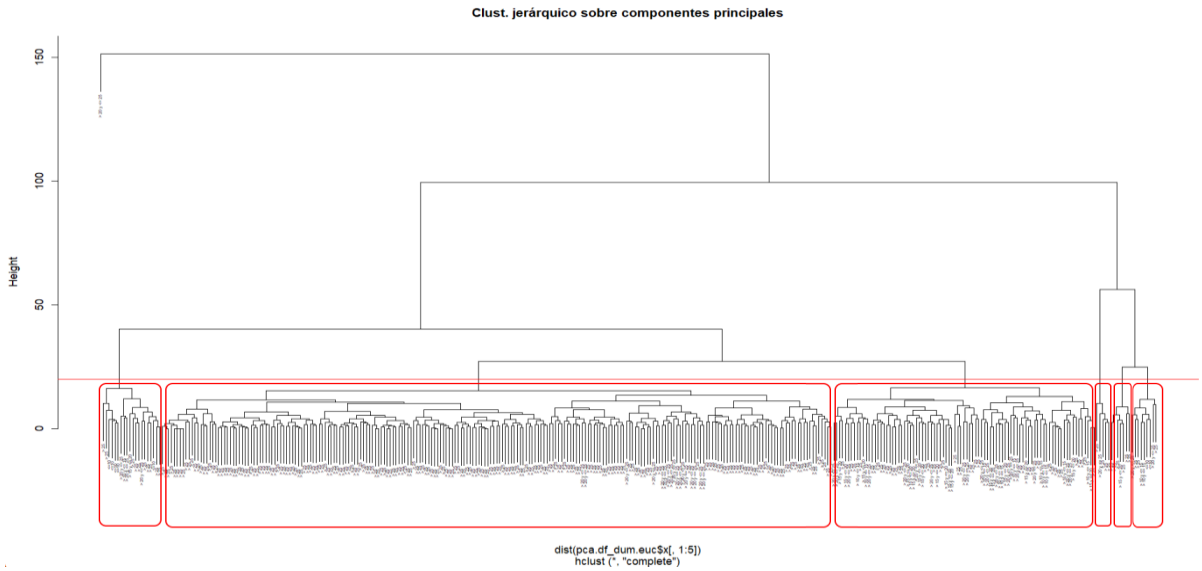


Figura 25 Dendograma de tres primeros componentes principales de conductores

K-means

Así también se genera un modelo *k-means* con K = 6 y 20 asignaciones aleatorias de clústeres iniciales, dado que al modelo ingresan 32 variables, al representar los grupos en dos dimensiones como se hace en la Figura 26 se está abstrayendo mucha información.

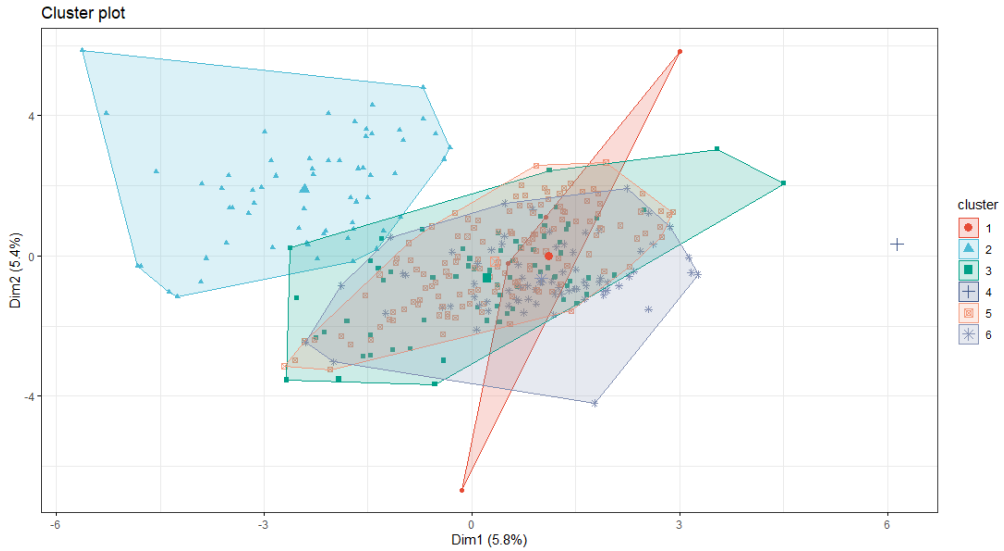


Figura 26 Clusters generados con *K-means*

De esta manera se presentan los resultados en cuanto a las infracciones con vehículos de la flota y uso de la misma en la Tabla 8.

Tabla 8 Resultados de agrupamiento jerárquico y *k-means* en variables de la flota

clusters_hc	N. cond.	Citaciones flota	Citaciones promedio flota	Puntos perdidos flota	Sancion	Promedio de Sancion flota	Promedio de puntos	Distancia promedio (km)	Tiempo promedio mov
1	13	49	3,8	0	$3.943	$80	18,04	13,26	6,87
2	3	5	1,7	14	$247	$49	18,00		
3	113	172	1,5	0	$13.287	$77	24,71	26,51	9,15
4	9	12	1,3	6	$586	$49	19,67	20,97	10,51
5	191	184	1,0	0	$13.116	$71	28,99	22,19	8,01
6	5	11	2,2	12	$720	$65	26,40		
Total	**334**	**433**	**1,3**	**32**	**$31.900**	**$74**	**26,73**	**23,74**	**8,63**

clusters_hc_pca	N. cond.	Citaciones flota	Citaciones promedio flota	Puntos perdidos flota	Sancion	Promedio de Sancion flota	Promedio de puntos	Distancia promedio (km)	Tiempo promedio mov
1	23	81	3,5	0	$6.829	$84	19,70	35,06	11,72
2	3	5	1,7	14	$247	$49	18,00		
3	97	131	1,4	0	$9.221	$70	24,58	22,71	8,06
4	9	12	1,3	6	$586	$49	19,67	20,97	10,51
5	197	193	1,0	0	$14.297	$74	29,07	22,20	8,08
6	5	11	2,2	12	$720	$65	26,40		
Total	**334**	**433**	**1,3**	**32**	**$31.900**	**$74**	**26,73**	**23,74**	**8,63**

clusters_km	N. cond.	Citaciones flota	Citaciones promedio flota	Puntos perdidos flota	Sancion	Promedio de Sancion flota	Promedio de puntos	Distancia promedio (km)	Tiempo promedio mov
1	4	9	2,3	17	$444	$49	19,63		
2	54	72	1,3	0	$5.928	$82	27,54	25,49	8,91
3	64	84	1,3	0	$6.312	$75	18,71	10,20	4,66
4	1	2	2,0	3	$42	$21	0,00		
5	139	228	1,6	0	$16.496	$72	29,53		
6	72	38	0,5	12	$2.678	$70	28,60	17,57	9,04
Total	**334**	**433**	**1,3**	**32**	**$31.900**	**$74**	**26,73**	**23,74**	**8,63**

En los grupos 3 y 5 de los modelos de agrupamiento jerárquico se concentra la mayoría de los conductores, así también las infracciones generadas con los vehículos de la flota; sin embargo, en el grupo 5 se registra un mayor puntaje seguido del grupo 6 y 3, así también el grupo 5 registra un menor número promedio de citaciones por conductor.

En relación al tiempo y distancia recorrida diaria promedio, si bien se observan mayores valores en ciertos grupos, es importante considerar la composición de cada grupo por tipo de vehículo.

El modelo *k-means* por su lado, presenta 4 grupos importantes; y, 2 con menor cantidad de conductores, aquí se evidencia que:

Los grupos 2, 5 y 6 contemplan a los conductores con mayor puntaje promedio.

Los grupos 2, 3, 5 y 6 registran un mayor monto en sanciones y sanciones promedio.

El grupo 1 presenta un mayor número de citaciones promedio por conductor.

El grupo 3 registra el menor puntaje promedio.

Considerando las infracciones de vehículos particulares en base a los grupos generados se tiene un comportamiento similar en los modelos de agrupamiento jerárquico en cuanto a las citaciones promedio y sanción, a excepción de los puntos perdidos, en la flota alcanzan 32, mientras en vehículos particulares llegan a 431,5, los cuales se concentran en los grupos 3 y 5; sin embargo, el mayor puntaje promedio perdido por conductor se registra en los grupos 4, 1 y 6, esto se registra en la Tabla 9.

Tabla 9 Resultados de agrupamiento jerárquico y *k-means* en variables de conductores

clusters_hc	N. cond.	N. citaciones	Citaciones promedio conductor	Puntos perdidos	Puntos perdidos promedio conductor	Sancion	Promedio de Sancion	Promedio de puntos
1	13	30	2,31	39,00	3,00	$2.345	$78	18,04
2	3							18,00
3	113	195	1,73	210,00	1,86	$15.617	$80	24,71
4	9	15	1,67	28,50	3,17	$761	$51	19,67
5	191	156	0,82	140,50	0,74	$15.258	$98	28,99
6	5	10	2,00	13,50	2,70	$866	$87	26,40
Total	**334**	**406**	**1,22**	**431,50**	**1,29**	**$34.847**	**$86**	**26,73**

clusters_hc_pca	N. cond.	N. citaciones	Citaciones promedio conductor	Puntos perdidos	Puntos perdidos promedio conductor	Sancion	Promedio de Sancion	Promedio de puntos
1	23	37	1,61	51,00	2,22	$2.753	$74	19,70
2	3							18,00
3	97	191	1,97	189,00	1,95	$15.732	$82	24,58
4	9	15	1,67	28,50	3,17	$761	$51	19,67
5	197	153	0,78	149,50	0,76	$14.735	$96	29,07
6	5	10	2,00	13,50	2,70	$866	$87	26,40
Total	**334**	**406**	**1,22**	**431,50**	**1,29**	**$34.847**	**$86**	**26,73**

clusters_km	N. cond.	N. citaciones	Citaciones promedio conductor	Puntos perdidos	Puntos perdidos promedio conductor	Sancion	Promedio de Sancion	Promedio de puntos
1	4	3	0,75	3,00	0,75	$105	$35	19,63
2	54	66	1,22	76,50	1,42	$4.284	$65	27,54
3	64	147	2,30	220,50	3,45	$11.290	$77	18,71
4	1							0,00
5	139	114	0,82	70,50	0,51	$10.724	$94	29,53
6	72	76	1,06	61,00	0,85	$8.444	$111	28,60
Total	**334**	**406**	**1,22**	**431,50**	**1,29**	**$34.847**	**$86**	**26,73**

Los puntos perdidos por infracciones de vehículos particulares de acuerdo al modelo *k-means* se concentran en el grupo 3 en donde también se registra el mayor número de puntos promedio perdidos y citaciones promedio por conductor.

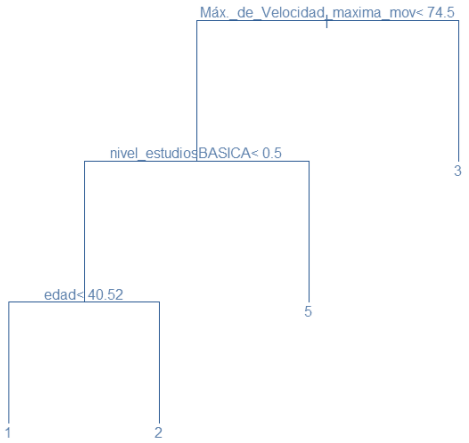
2.2.1.3.2. Modelos supervisados

Para la generación de los modelos supervisados utilizamos como variable de etiqueta o control el resultado del modelo *k-means*, se generan árboles de decisión y SVM con los datos originales y con datos balanceados por sobremuestreo, considerando un grupo de entrenamiento y de prueba, con dos corridas, la primera (60% - 40%) y segunda (70% - 30%). Para un mayor control en la generación de modelos se establecen como parámetros “trainControl” para todos los modelos, 10 iteraciones de remuestreo y cinco validaciones cruzadas (method = "repeatedcv", number = 10 y repeats = 5).

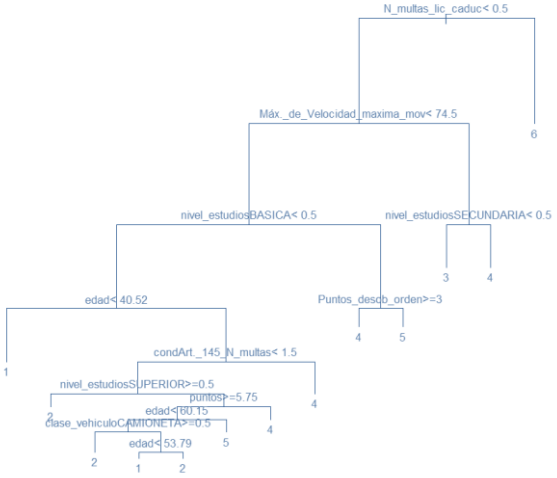
Árbol de decisión

Como parámetro de los modelos de árboles de decisión se utiliza un mínimo de observaciones por nodo de 10, los árboles generados se presentan en la Figura 27.

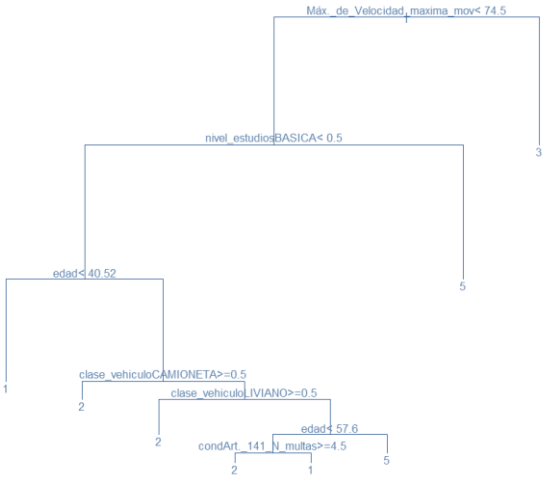
Árbol de clasificación (60%-40%)



Árbol de clasificación balanceado (60%-40%)



Árbol de clasificación (70%-30%)



Árbol de clasificación balanceado (70%-30%)

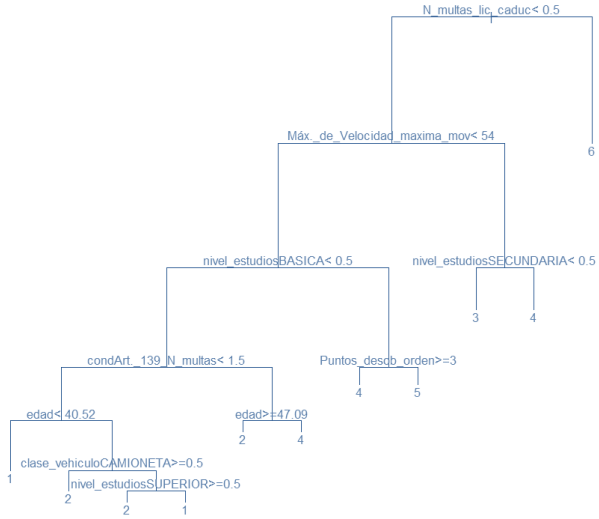


Figura 27 Árboles de decisión con y sin balanceo (60%-40%) y (70%-30%)

Máquina de soporte vectorial SVM

Como parámetros de los modelos de máquinas de soporte vectorial se utiliza el *kernel* radial, una función sigma = (0.001, 0.01, 0.1, 0.5, 1) y una función de costo C = (1, 20, 50, 100, 200, 500, 700), la evolución de la exactitud de los modelos se presenta en la Figura 28.

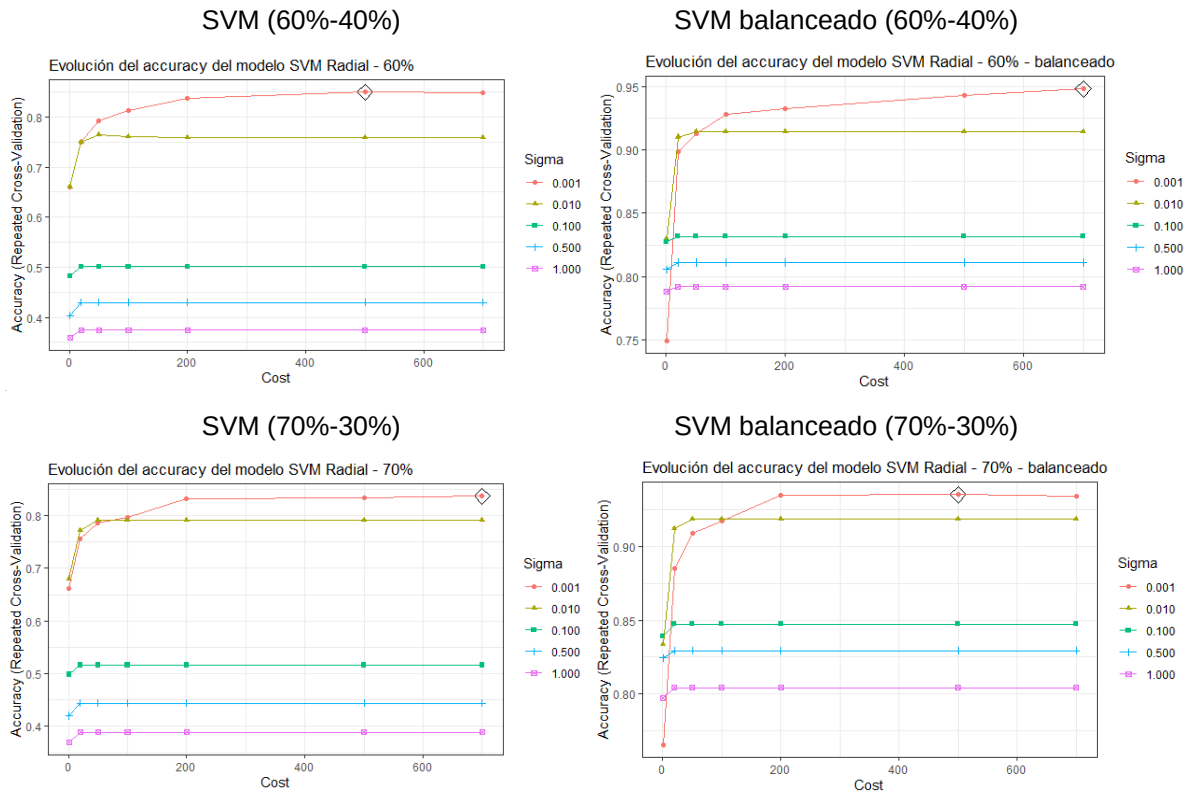


Figura 28 SVM con y sin balanceo (60%-40%) y (70%-30%)

2.2.1.4. Evaluación

Para la evaluación de los modelos, se generan matrices de confusión, se calcula la exactitud de cada modelo, y la precisión, sensibilidad y *F1-score* por clase.

Los resultados de los modelos con el 70% de datos como grupo de entrenamiento y 30% de prueba se muestran en la Tabla 10.

Tabla 10 Métricas de árboles de decisión y SVM con 70%-30%

Entrenamiento	70%			Prueba		30%				
		grupo de prueba								
	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
árbol	1	20	5	1	2	3	0	65%	83%	73%	80,4%
	2	4	27	0	2	1	0	79%	82%	81%	
	3	0	0	20	0	1	0	95%	95%	95%	
	4	0	0	0	0	0	0		0%		
	5	0	1	0	1	19	0	90%	79%	84%	
	6	0	0	0	0	0	0				
		grupo de prueba									
árbol balanceado	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
	1	20	11	0	1	6	0	53%	83%	65%	71,0%
	2	3	20	0	2	1	0	77%	61%	68%	
	3	1	0	20	0	2	0	87%	95%	91%	
	4	0	2	1	1	0	0	25%	20%	22%	
	5	0	0	0	1	15	0	94%	63%	75%	
	6	0	0	0	0	0	0				
		grupo de prueba									
svm	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
	1	23	1	0	0	1	0	92%	96%	94%	89,7%
	2	0	30	2	2	1	0	86%	91%	88%	
	3	1	0	19	0	1	0	90%	90%	90%	
	4	0	0	0	3	0	0	100%	60%	75%	
	5	0	2	0	0	21	0	91%	88%	89%	
	6	0	0	0	0	0	0				
		grupo de prueba									
svm balanceado	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
	1	23	1	0	0	2	0	88%	96%	92%	88,8%
	2	0	30	2	2	1	0	86%	91%	88%	
	3	1	0	19	0	1	0	90%	90%	90%	
	4	0	0	0	3	0	0	100%	60%	75%	
	5	0	2	0	0	20	0	91%	83%	87%	
	6	0	0	0	0	0	0				

Los modelos de SVM y SVM balanceado con entrenamiento 70%-30% presentan una mayor exactitud, 89.7% y 88.8% respectivamente.

Los resultados de los modelos con el 60% de datos como grupo de entrenamiento y 40% de prueba se muestran en la Tabla 11.

Tabla 11 Métricas de árboles de decisión y SVM con 60%-40%

Entrenamiento	60%			Prueba	40%					
		grupo de prueba								
	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
árbol	1	19	3	0	1	0	0	83%	59%	69%	76,2%
	2	11	42	1	4	9	0	63%	93%	75%	
	3	0	0	25	0	2	0	93%	96%	94%	
	4	0	0	0	0	0	0		0%		
	5	2	0	0	1	23	0	88%	68%	77%	
	6	0	0	0	0	0	0				
		grupo de prueba								
árbol balanceado	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
	1	26	11	1	1	1	0	65%	81%	72%	79,7%
	2	4	34	0	1	4	0	79%	76%	77%	
	3	0	0	24	0	2	0	92%	92%	92%	
	4	0	0	1	3	0	0	75%	50%	60%	
	5	2	0	0	1	27	0	90%	79%	84%	
	6	0	0	0	0	0	0				
		grupo de prueba								
svm	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
	1	27	4	0	1	2	0	79%	84%	82%	85,3%
	2	1	39	2	1	2	0	87%	87%	87%	
	3	1	1	24	0	2	0	86%	92%	89%	
	4	0	0	0	4	0	0	100%	67%	80%	
	5	3	1	0	0	28	0	88%	82%	85%	
	6	0	0	0	0	0	0				
		grupo de prueba								
svm balanceado	predicciones	1	2	3	4	5	6	Precisión	Sensibilidad	F1-score	Exactitud
	1	28	4	0	0	1	0	85%	88%	86%	86,0%
	2	1	39	2	3	2	0	83%	87%	85%	
	3	1	1	24	0	2	0	86%	92%	89%	
	4	0	0	0	3	0	0	100%	50%	67%	
	5	2	1	0	0	29	0	91%	85%	88%	
	6	0	0	0	0	0	0				

Dado que los datos no están balanceados; y, la diferencia en los resultados con y sin balance es apenas del 0.9%, se selecciona como mejor modelo al SVM balanceado con entrenamiento del 70%-30%.

El modelo seleccionado genera similares grupos a los generados con el modelo *K-means*. Con ello se podrá clasificar a nuevos conductores, siendo útil también cuando se requiera una actualización de la clasificación.

2.3. SISTEMA DE INFORMACIÓN

Para el desarrollo del tablero se utilizó Microsoft Power BI, en donde se integran los datos obtenidos de los procesos de *web scraping* y los resultados del modelo de clasificación de conductores para tener una visión general de la flota, su uso; y, relacionarla con la caracterización de conductores, para de esta manera facilitar la generación de análisis adicionales, el modelado de datos se visualiza en la Figura 29.

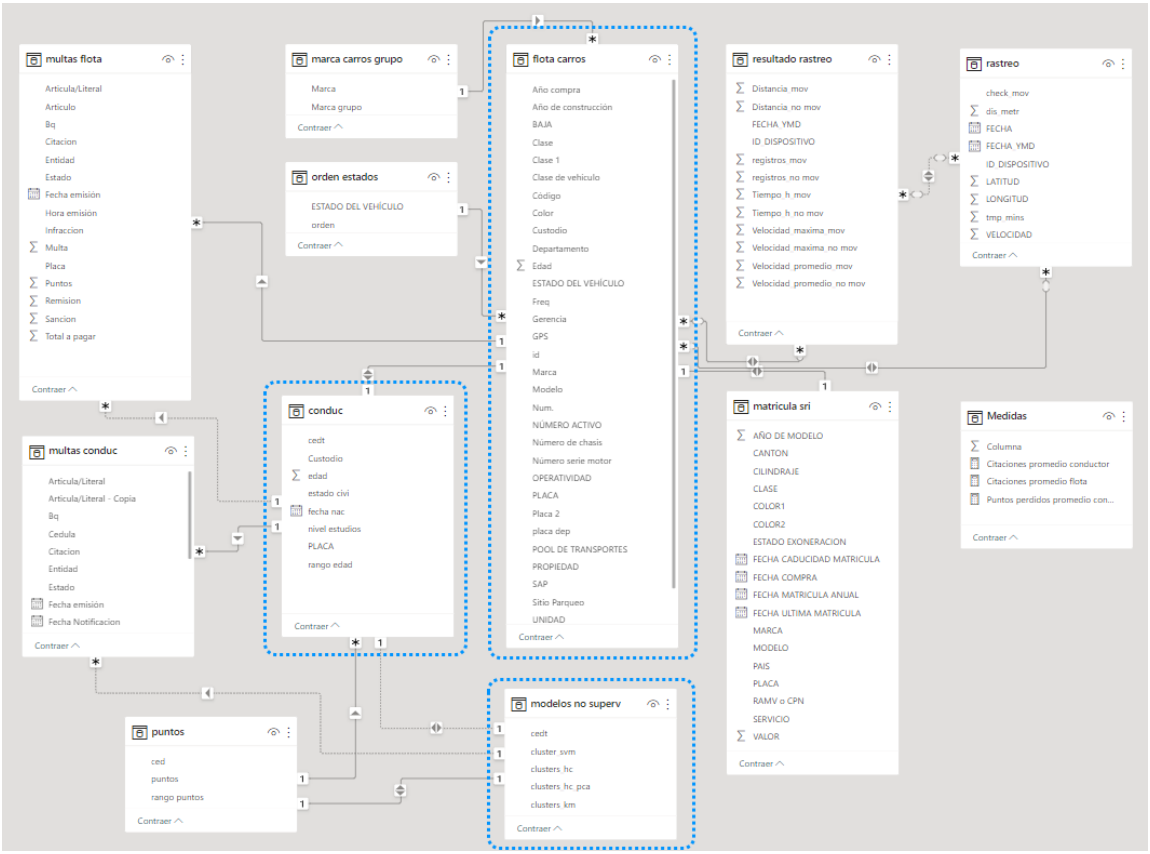


Figura 29 Modelo de datos

2.3.1. Flota

En la página Flota presentada en la Figura 30 se puede cuantificar a los vehículos por clase, marca, estado, edad, si cuenta con servicio de rastreo y gerencia asignada.

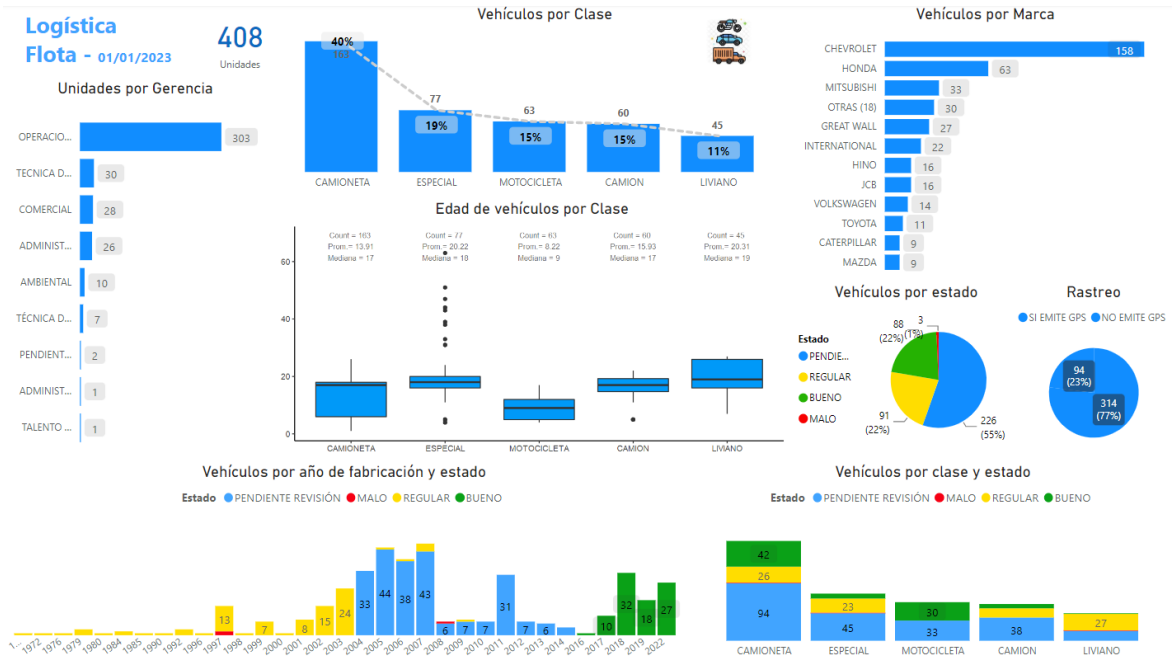


Figura 30 Flota vehicular

2.3.2. Monitoreo

En esta sección se puede analizar tiempos de uso promedio diario de los vehículos, su distancia recorrida, de manera general y por clase de vehículos, gerencias, entre otros; se visualiza además la evolución de velocidad promedio, velocidad máxima, tiempo de uso y distancias promedio, se presenta una captura en la Figura 31.

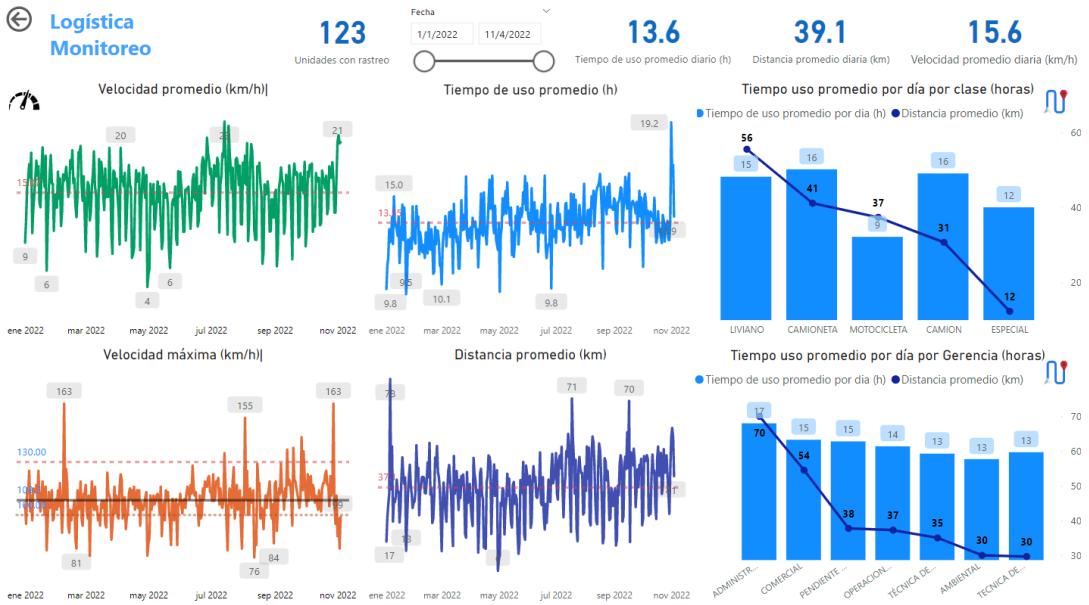


Figura 31 Monitoreo de flota vehicular -1

Como extensión a la página anterior, se pueden visualizar el tiempo de uso promedio por departamento y unidad, así también la evolución de dichos indicadores por departamento en la Figura 32.

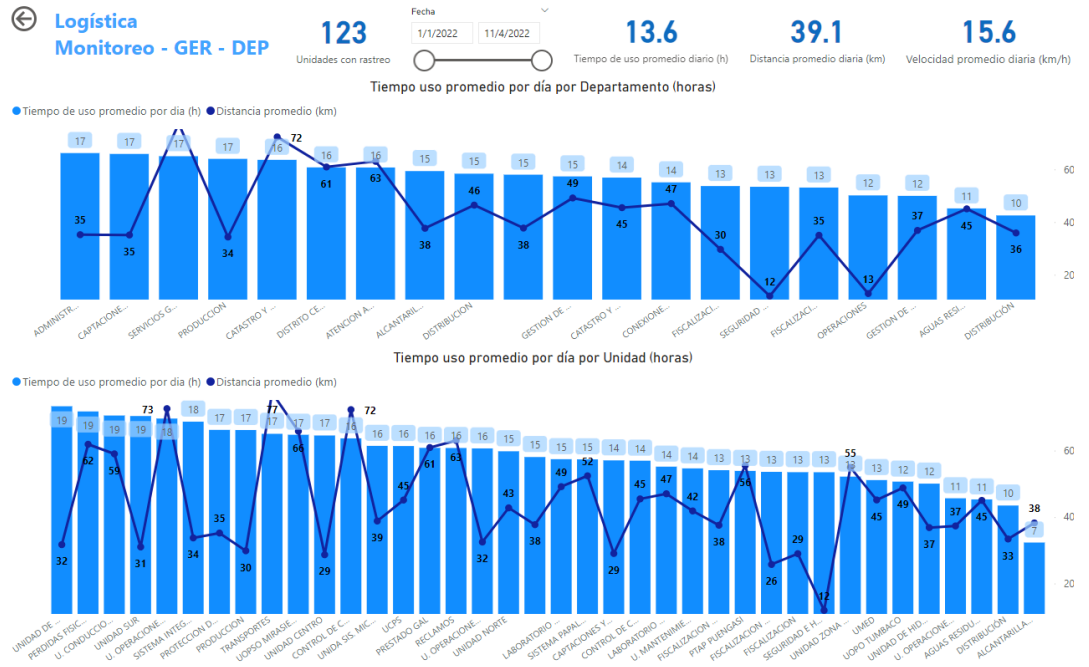


Figura 32 Monitoreo de flota vehicular – 2

2.3.3. Infracciones con vehículos de la flota - modelos

En esta página se presentan las infracciones de tránsito generadas con los vehículos de la Empresa, se cuantifican los mismos y la sanción económica correspondiente por distintas variables como estado, tipo de multa, clase de vehículo, gerencia; se observa: la evolución de las mismas, y el ranking de vehículos con mayor cantidad de infracciones. Esto se puede observar en la Figura 33, además, se incluyen los grupos generados a partir de los modelos para observar el comportamiento de cada uno.

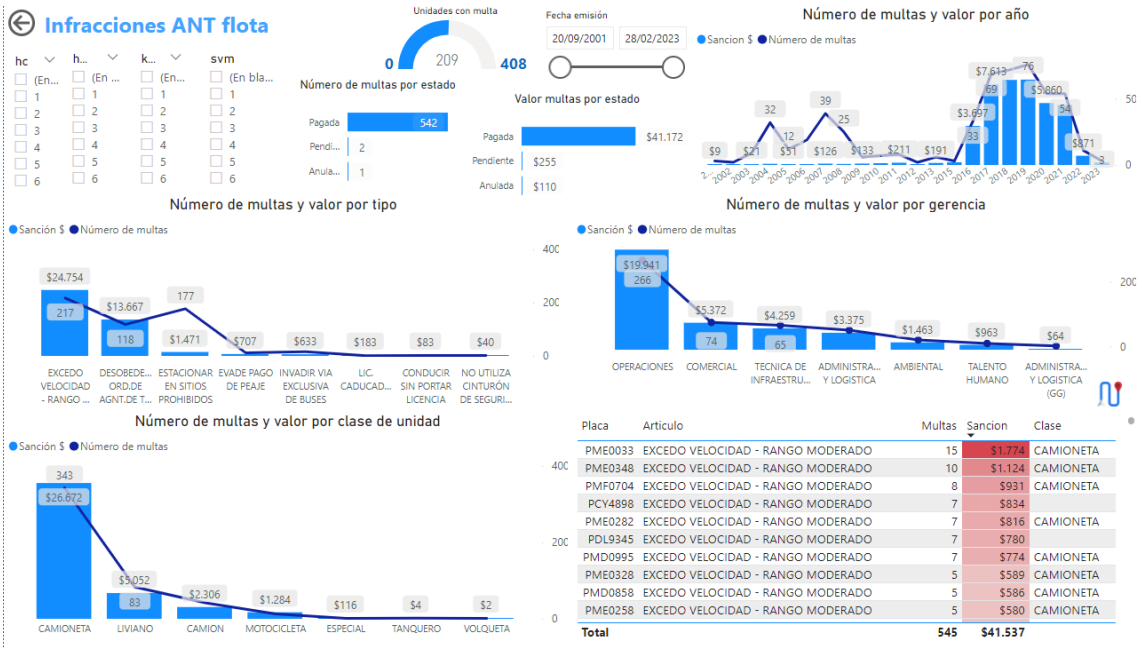
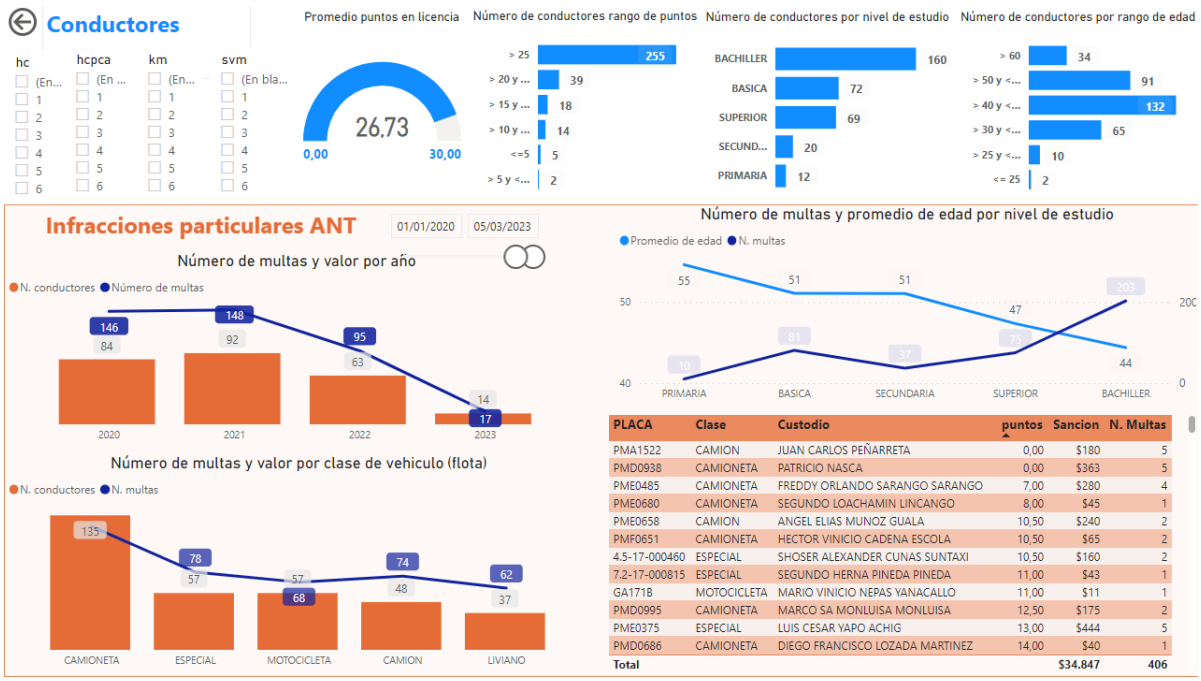


Figura 33 Infracciones flota vehicular

2.3.4. Conductores e infracciones con cualquier vehículo - modelos

En la Figura 34 se observa una captura de la página de conductores, en donde se presenta la distribución de los conductores en los grupos creados con los modelos, se identifican a los conductores por rangos de puntos de conducción, rangos de edad y nivel de estudio, además se muestran las infracciones de tránsito generadas por los conductores con cualquier vehículo, se cuantifican las mismas por clase de vehículo y se observa su evolución, así también el ranking de conductores con menor cantidad de puntos y mayor cantidad de infracciones.

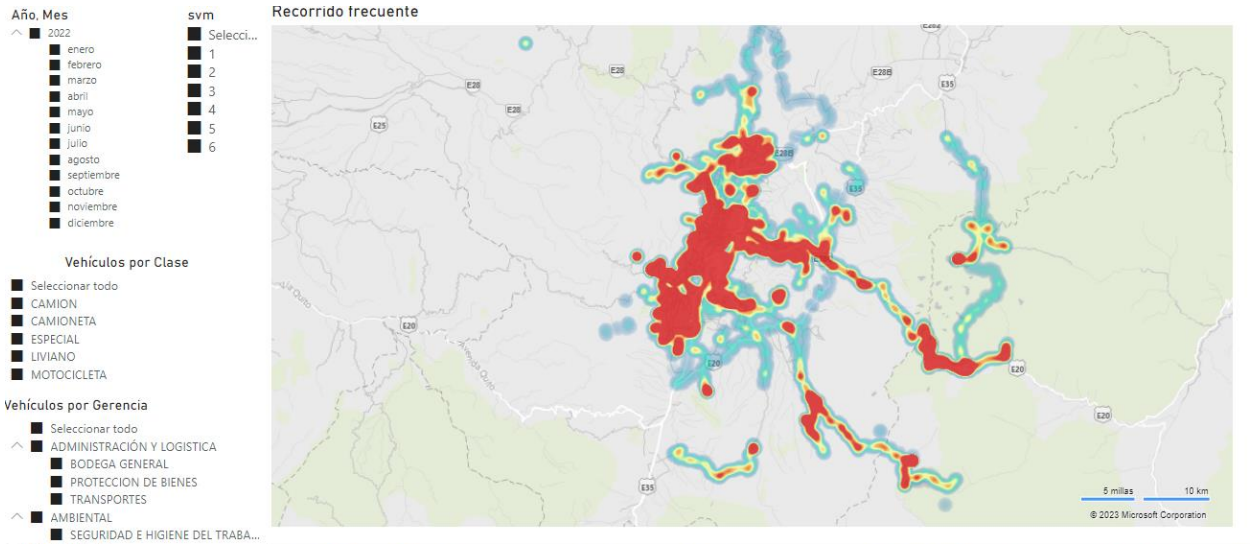


PLACA	Clase	Custodio	puntos	Sancion	N. Multas
PMA1522	CAMION	JUAN CARLOS PEÑARRETA	0,00	$180	5
PMD0938	CAMIONETA	PATRICIO NASCA	0,00	$363	5
PME0485	CAMIONETA	FREDDY ORLANDO SARANGO SARANGO	7,00	$280	4
PME0680	CAMIONETA	SEGUNDO LOACHAMIN LINCANGO	8,00	$45	1
PME0658	CAMION	ANGEL ELIAS MUNOZ GUALA	10,50	$240	2
PMF0651	CAMIONETA	HECTOR VINICIO CADENA ESCOLA	10,50	$65	2
4.5-17-000460	ESPECIAL	SHOSER ALEXANDER CUNAS SUNTAXI	10,50	$160	2
7.2-17-000815	ESPECIAL	SEGUNDO HERNA PINEDA PINEDA	11,00	$43	1
GA1718	MOTOCICLETA	MARIO VINICIO NEPAS YANACALLO	11,00	$11	1
PMD0995	CAMIONETA	MARCO SA MONLUISA MONLUISA	12,50	$175	2
PME0375	ESPECIAL	LUIS CESAR YAPO ACHIG	13,00	$444	5
PMD0686	CAMIONETA	DIEGO FRANCISCO LOZADA MARTINEZ	14,00	$40	1
Total				$34.847	406

Figura 34 Conductores e infracciones vehículos particulares - modelos

2.3.5. Rutas frecuentes - Visualización en mapa

La Figura 35 permite identificar el recorrido frecuente de los vehículos de la flota por clase, gerencia y grupos creados a partir del modelo, se observa además un resumen de los principales indicadores ligados a las infracciones de tránsito de la flota.



cluster_svm	N. cond.	Citaciones flota	Citaciones promedio flota	Puntos perdidos flota	Sancion	Promedio de Sancion flota	Promedio de puntos	Distancia promedio (km)	Tiempo promedio mov
1	14	3	0,2	0	$182	$61	27,45	19,02	9,53
2	9	13	1,4	0	$993	$76	25,64	26,36	11,67
3	27	40	1,5	0	$3.708	$93	27,75	31,14	10,79
4	1						11,63	32,84	8,86
5	15	8	0,5	0	$444	$56	28,62	18,43	9,54
6							24,00		
Total	66	64	1,0	0	$5.327	$83	26,80	24,52	10,30

Figura 35 Recorrido frecuente de vehículos por clase, gerencia y grupo

La Gerencia de Administración y Logística registra mayor presencia en el Edificio Matriz y las Bodegas Generales.
La Gerencia de Ambiente registra una mayor presencia en las Oficinas del Departamento de Seguridad e Higiene del Trabajo.
La Gerencia Comercial registra mayor presencia en las Oficinas de la Unidad de Laboratorio de Medidores y el Edificio Matriz.
La Gerencia de Operaciones registra una mayor dispersión en sus recorridos, al ser el área encargada de la mayor cantidad de procesos operativos enfocados al cumplimiento de la misión de la Empresa.
La Gerencia Técnica de Infraestructura registra mayor presencia en el Edificio Matriz.

En el tablero desarrollado se puede apreciar esto a mayor detalle y revisar incluso por clase de vehículo, ya que debido a su naturaleza registran distintos recorridos.

3. RESULTADOS

Para un correcto control de la flota vehicular de la Empresa es imprescindible contar con un panorama de la flota y de sus conductores, esto es, tener claridad del número de vehículos por clase, año de fabricación, estado, entre otros; conocer las características de los conductores, su edad, formación, puntos en su licencia de conducción, etc.

Además, es necesario contar con información del uso de los vehículos de la flota para identificar posibles problemas de subutilización.

De esta manera, se realiza el análisis descriptivo de la flota de vehículos de la Empresa, características y comportamiento de los conductores durante la conducción, y uso de los vehículos por gerencias y departamentos para generar recomendaciones de optimización de la flota.

3.1. Análisis de la flota vehicular

Básicamente lo que se pretende con este análisis es como primer punto, dimensionar la flota vehicular de la Empresa Pública; y segundo, caracterizar la flota en base al detalle de los vehículos.

La flota vehicular de la Empresa Pública de acuerdo a la clase de vehículos se puede observar en la Figura 36, está conformada por 408 vehículos, de los cuales el 40% son camionetas, 19% corresponde a maquinaria pesada, 15% motocicletas, 15% camiones y el restante 11% a vehículos livianos.

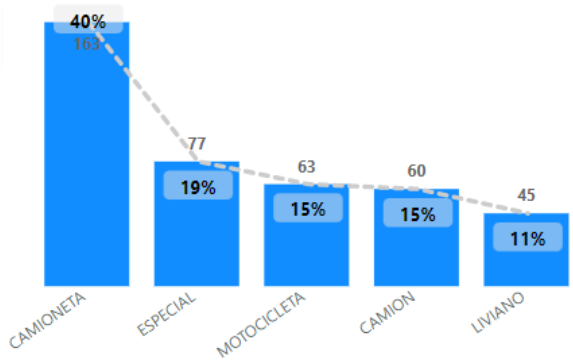


Figura 36 Vehículos por clase

En la Figura 37 se evidencia que las marcas con mayor presencia en la flota vehicular de la Empresa son Chevrolet, Honda y Mitsubishi, las cuales concentran el 62%.

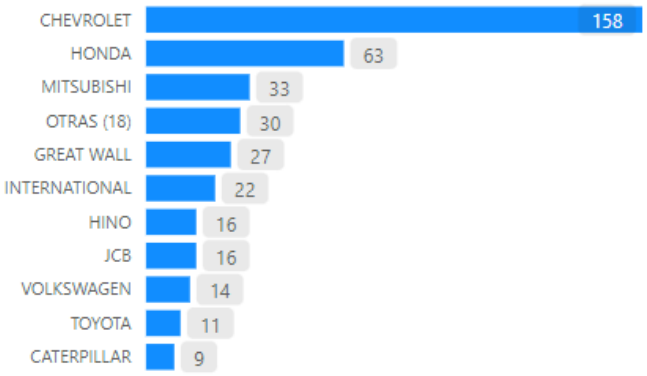


Figura 37 Vehículos por marca

Cabe señalar que todas las motocicletas son Honda, a diferencia de Chevrolet que está distribuido el 80% en camionetas, 16% en vehículos livianos y 3% en camiones, Mitsubishi por su parte tiene el 79% en camiones y 21% en livianos.

En 2022 se adquirieron 27 camionetas Greatwall, el resto de marcas no presentan una concentración en una determinada clase.

La edad promedio de la flota vehicular es de 15 años, siendo los vehículos livianos y especiales aquellos con mayor edad, superando los 20 años. Las motocicletas por su parte registran una edad promedio de 8 años, esto se observa en la Figura 38.

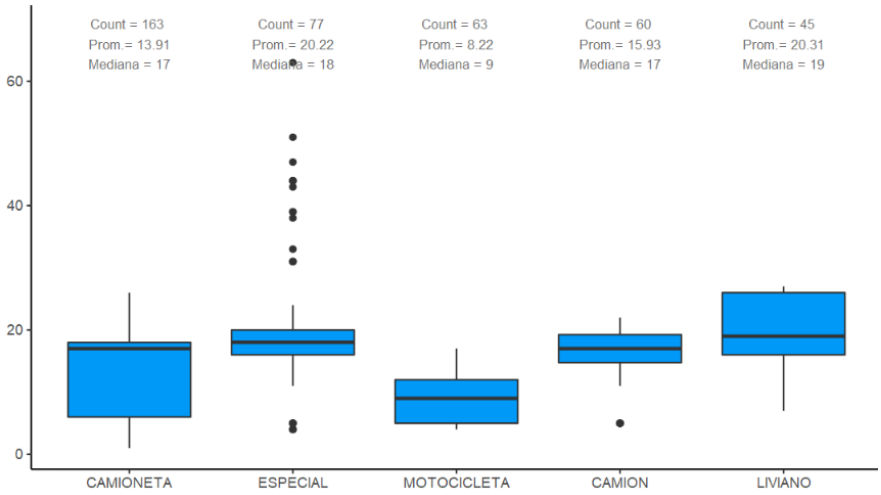


Figura 38 Edad de vehículos por clase

La Figura 39 muestra que el 55% de los vehículos se encuentran pendientes de revisión, el 22% están en buen estado, 22% en estado regular y el 1% se encuentra en mal estado.

El 63% de los camiones se encuentran pendientes de revisión, en las camionetas y vehículos especiales están pendientes el 58%, en las motocicletas este valor alcanza el 52% y para los vehículos livianos el 36%.

Se encuentran en estado regular el 60% de los vehículos livianos, el 29% de los vehículos especiales, el 25% de camiones y el 16% de camionetas.

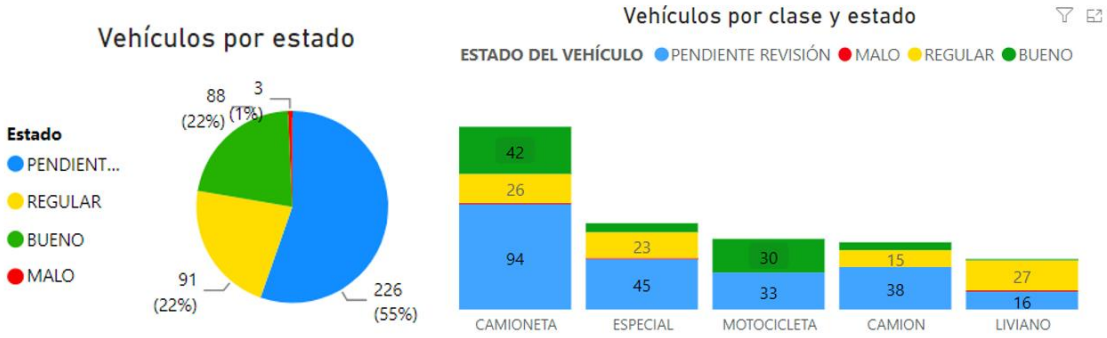


Figura 39 Vehículos por estado y clase

Se observa una mayor participación de vehículos en buen estado en motocicletas (48%) y camionetas (26%). Existe una camioneta, un vehículo especial y un liviano en mal estado.
En la Figura 40 se puede identificar que los vehículos de 2016 en adelante se encuentran en buen estado, mientras aquellos de años inferiores a 2003 se encuentran en estado regular; la mayor cantidad de vehículos de 2004 a 2014 tienen pendiente la revisión para determinar su estado, y dado que estos vehículos representan el 55% de la flota, es necesaria una pronta evaluación.

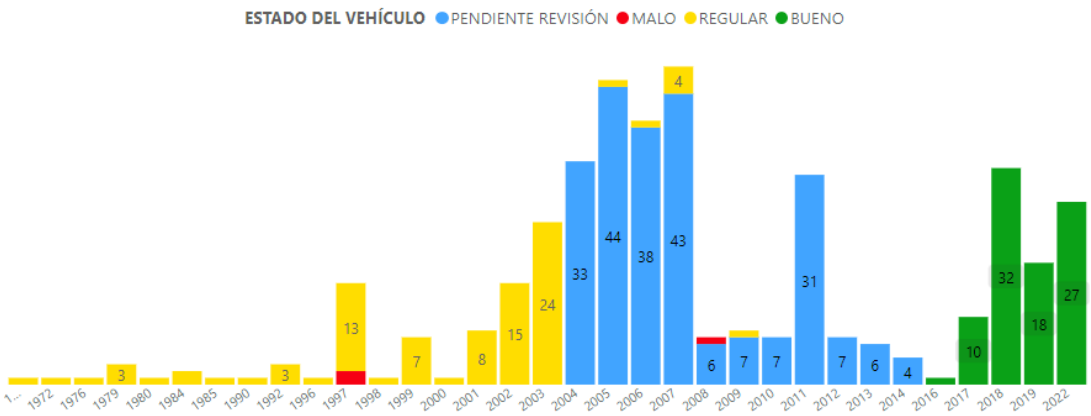


Figura 40 Vehículos por año de fabricación y estado

En relación al servicio de monitoreo satelital, de acuerdo a los registros recibidos el 77% de vehículos de la flota tienen activo el servicio, esto se registra en la Figura 41.

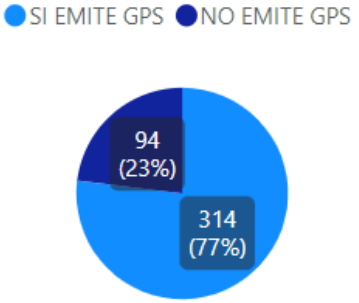


Figura 41 Vehículos por servicio de monitoreo satelital

3.2. Análisis de conductores

La Figura 42 muestra que los conductores de la flota vehicular de la Empresa Pública registran en promedio 26.73 puntos en sus licencias de conducir, con una mayor participación en el rango de mayores a 25 puntos (77%), seguidos de 20 a 25 puntos (12%), y de 15 a 20 puntos (5%). Cabe señalar que, existen siete conductores con menos de 10 puntos en su licencia de conducción.

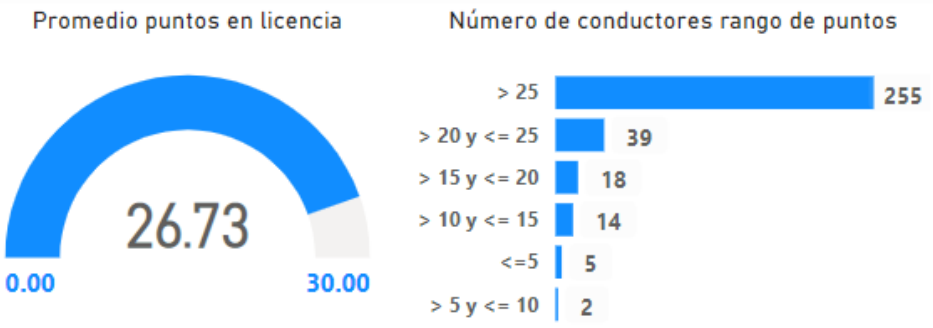


Figura 42 Promedio de puntos y distribución por rango de puntos

En la Figura 43 se visualiza que existe una mayor cantidad de conductores con nivel de estudio bachiller (48%), seguido de educación básica (22%) y superior (21%). En relación a su edad, existe una mayor cantidad de conductores con edades entre 40 y 50 años (40%), seguidos de conductores entre 50 y 60 años (27%); y, conductores entre 30 y 40 años (20%).

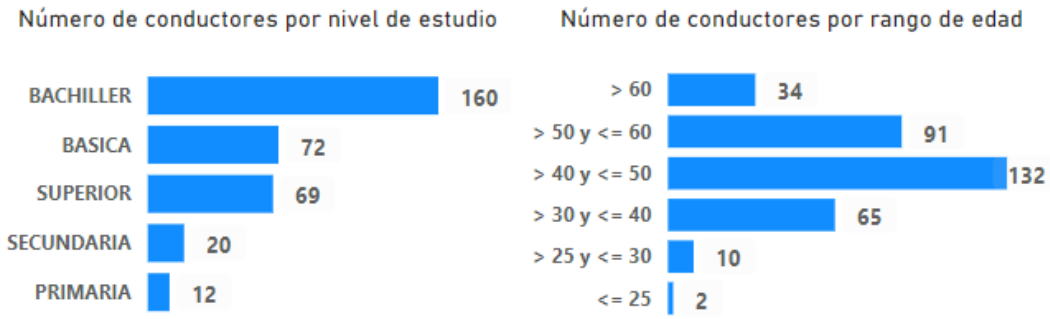


Figura 43 Distribución de conductores por nivel de estudio y rango de edad

3.3. Análisis de uso de vehículos

De acuerdo a los datos señalados previamente el 77% de la flota tiene el servicio de monitoreo satelital activo; sin embargo, al momento de analizar los datos de dicho servicio, se observa como lo muestra la Figura 44, que de enero a noviembre de 2022 únicamente 123 vehículos reportan su ubicación, por cuanto el análisis se realiza sobre dichos vehículos. Se tiene un tiempo de uso promedio de los vehículos de 13.6 horas, con una distancia promedio diaria de 39 km y una velocidad promedio diaria de 15.6 km/h.

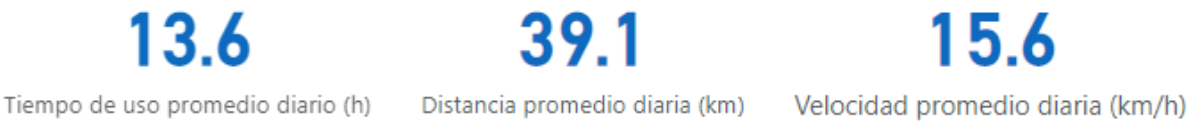


Figura 44 Indicadores de uso de vehículos

En los camiones y camionetas se registra un mayor tiempo de uso promedio por día, con un total de 16 horas, seguido de los vehículos livianos con 15 horas, vehículos especiales con 12 horas y motocicletas con un uso de 9 horas, esto se presenta en la Figura 45 en conjunto con las distancias recorridas promedio.

Se registran mayores distancias recorridas en vehículos livianos con un promedio diario de 56km, seguidos de camionetas con 41km y motocicletas con 37km, los camiones recorren en promedio 31km, mientras los vehículos especiales alcanzan 12km diarios en promedio.

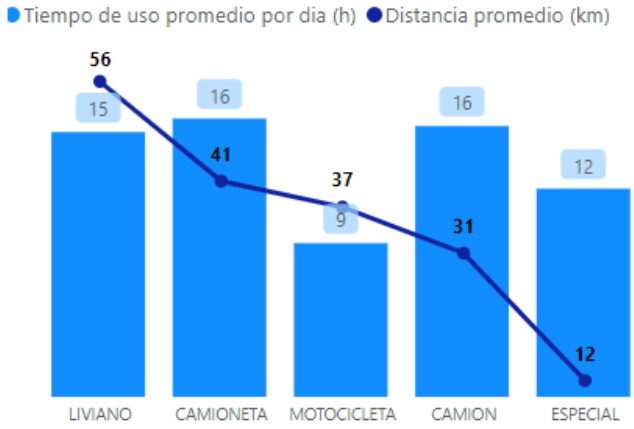


Figura 45 Tiempo de uso promedio por día por clase (h)

También se pueden revisar estas variables por gerencia asignada, como se visualiza en la Figura 46, es así que la Gerencia de Administración y Logística registra un mayor tiempo de uso y distancia promedio diaria, con 17 horas y 70km respectivamente, seguida de la Gerencia Comercial, la cual registra un promedio diario de uso de los vehículos de 15 horas y una distancia promedio de 54km, si bien las restantes gerencias tienen un tiempo de uso promedio de entre 13 y 15 horas, la distancia recorrida promedio por día está entre 30km y 38km únicamente.

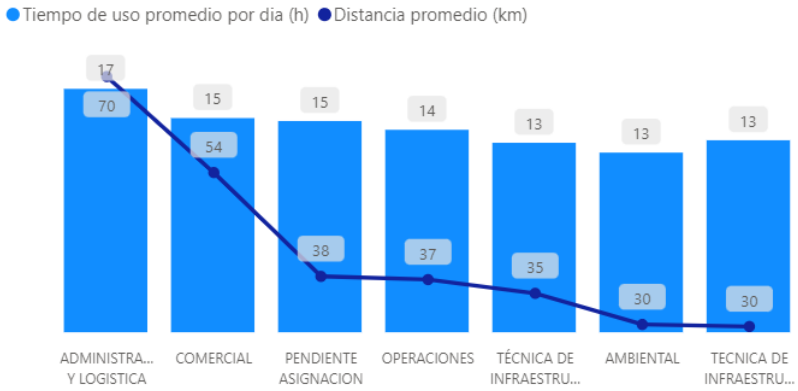


Figura 46 Tiempo de uso promedio por día por Gerencia

La distancia promedio diaria de la flota presenta ligeros incrementos en días específicos de los meses de enero, julio y septiembre; a excepción de esto, el comportamiento es estable durante todo el año, se puede visualizar en la Figura 47.

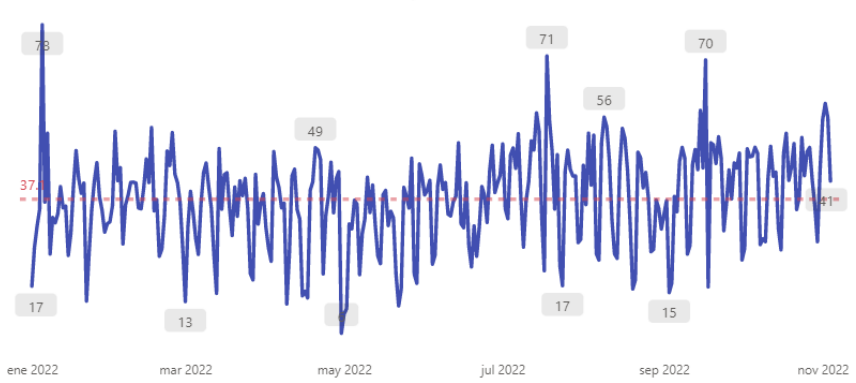


Figura 47 Distancia promedio (km)

El tiempo de uso promedio de los vehículos presenta un ligero incremento en el segundo semestre del año, con días específicos de mayor uso en noviembre, mayo y junio, se presenta de manera gráfica en la Figura 48.

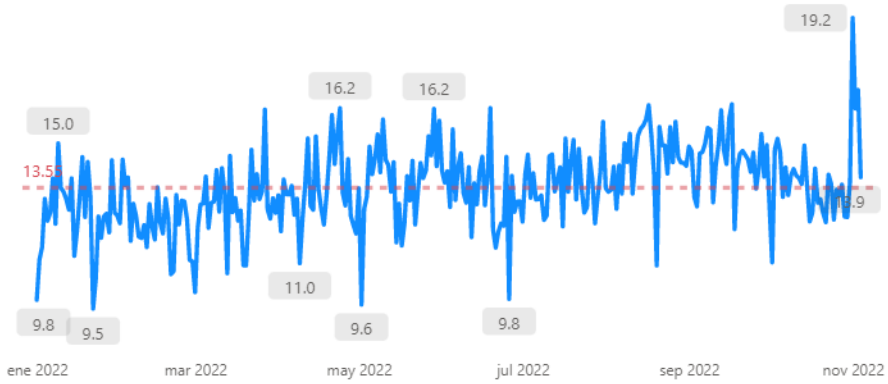


Figura 48 Tiempo de uso promedio (h)

La Figura 49 muestra que la velocidad promedio de la flota no supera los 30km/h durante todo el año. Para vehículos livianos alcanza los 50km/h.

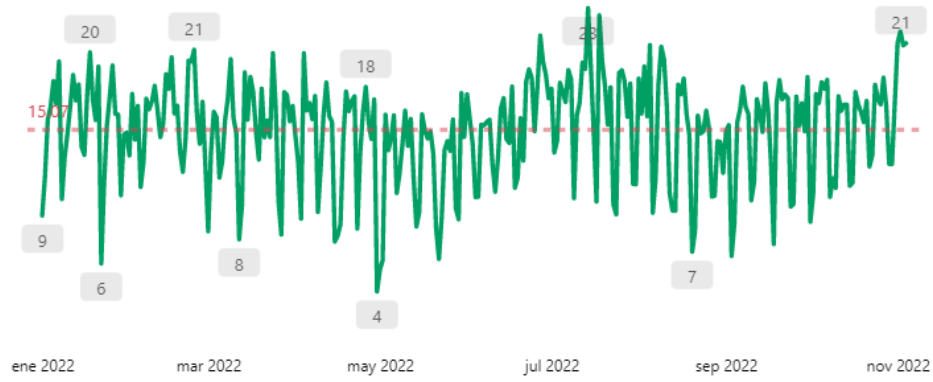


Figura 49 Velocidad promedio

La velocidad máxima en cambio alcanza valores de hasta 163km/h, con una gran cantidad de excesos de velocidad (superior a 100km/h), un detalle a considerar es que la velocidad máxima superior a los 100km/h en vehículos livianos no es tan frecuente comparada con las camionetas, esto se visualiza a mayor detalle en la Figura 50.

Los conductores de los vehículos PMA7855, PMA7848, PMA7857, GA152L, PMA3107, PMA7872 y PMA7849 registran velocidades de conducción fuera del rango moderado (mayor a 130km/h).

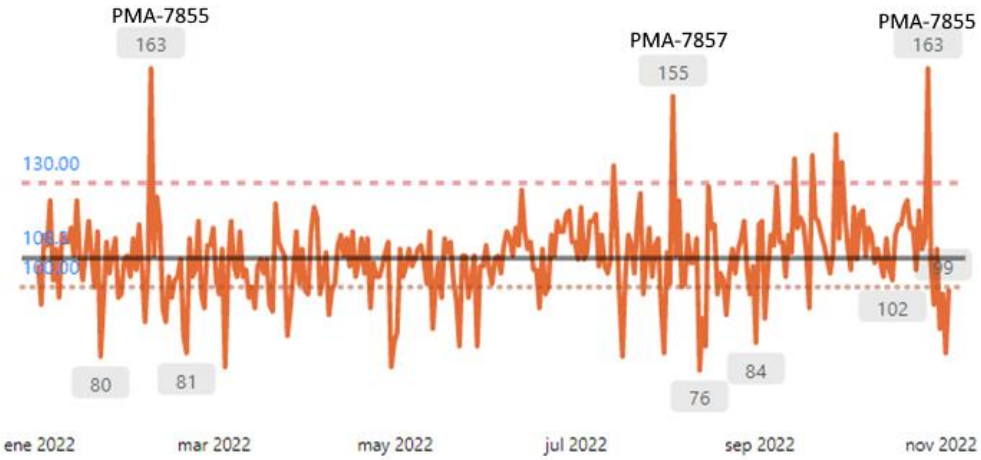


Figura 50 Velocidad máxima (km/h)

La flota vehicular de la Empresa Pública está asignada a varias gerencias, departamentos y unidades, en base a sus necesidades, la Figura 51 permite visualizar el tiempo de uso y distancia promedios por departamentos, en donde sobresalen por una posible subutilización el Departamento de Seguridad e Higiene del Trabajo y algunos departamentos de la Gerencia de Operaciones por registrar una menor distancia recorrida promedio, así también, el Departamento de Alcantarillado presenta un menor uso de vehículos, seguido de este se encuentra el Departamento de Distribución, Aguas Residuales y Operaciones Centro; lo propio sucede con el Departamento de Seguridad e Higiene del Trabajo, el cual registra una baja distancia recorrida promedio.

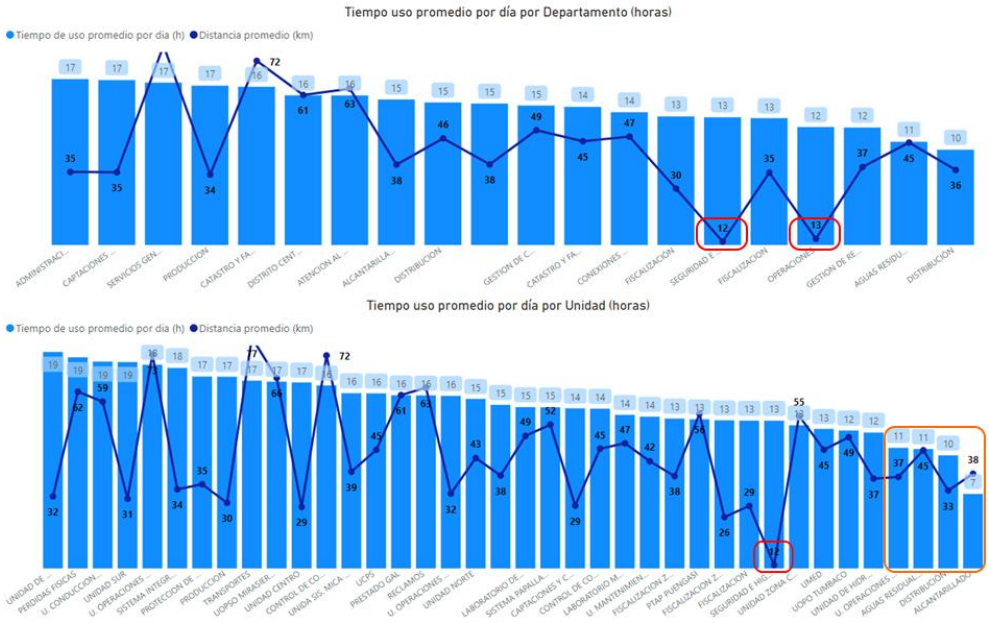


Figura 51 Tiempo de uso y distancia promedios diarios por Departamento y Unidad

3.4. Análisis de infracciones - vehículos

Las infracciones de tránsito toman relevancia en este análisis ya que conllevan una multa económica importante, en este sentido, es clave identificar a los mayores infractores y las infracciones más comunes para evitarlas, este detalle se presenta en la Figura 52.

Se registran un total de 545 infracciones desde 2001 al 28 de febrero de 2023, las cuales corresponden a 209 vehículos, dichas infracciones generaron un total en multas económicas de más de $41 mil dólares, en su mayoría ya pagados.

Figura 52 Infracciones de tránsito ANT generadas por vehículos de la flota

Las infracciones de tránsito generadas por los vehículos de la flota se concentran en los años posteriores a 2015, siendo 2017, 2018 y 2019 aquellos con mayor cantidad de infracciones, de manera gráfica se puede apreciar en la Figura 53.

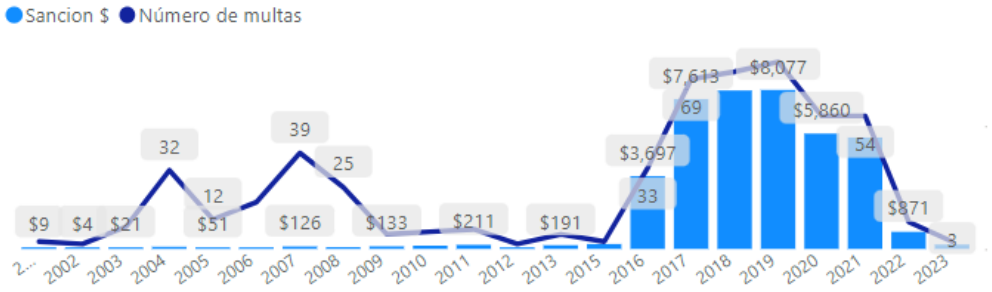


Figura 53 Número de infracciones y multa económica por año

De acuerdo a la Figura 54, las principales infracciones se dan por exceso de velocidad, estacionar en lugares prohibidos y desobedecer a un agente de tránsito, ésta última conlleva una multa económica mayor, por ende, en cuanto a monto se sitúa como el segundo tipo de infracción.

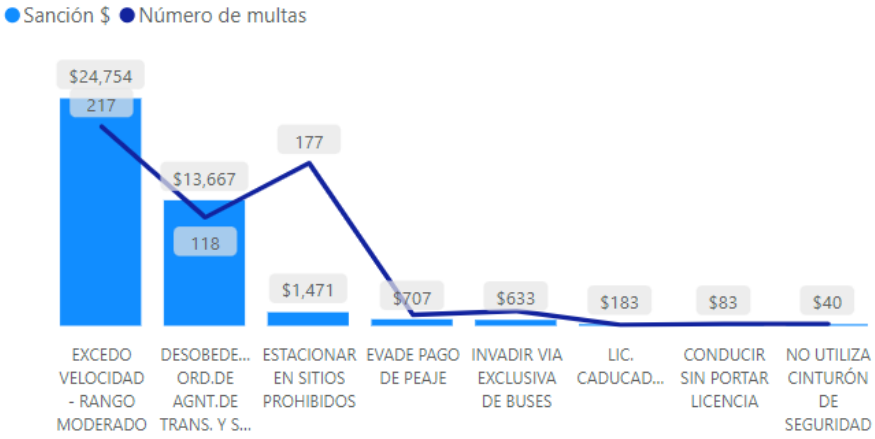


Figura 54 Número de infracciones y multa económica por tipo

Como se muestra en la Figura 55, el 63% de las infracciones se concentran en las camionetas, seguidas de éstas se encuentran los vehículos livianos con el 13% y los camiones con el 8%.

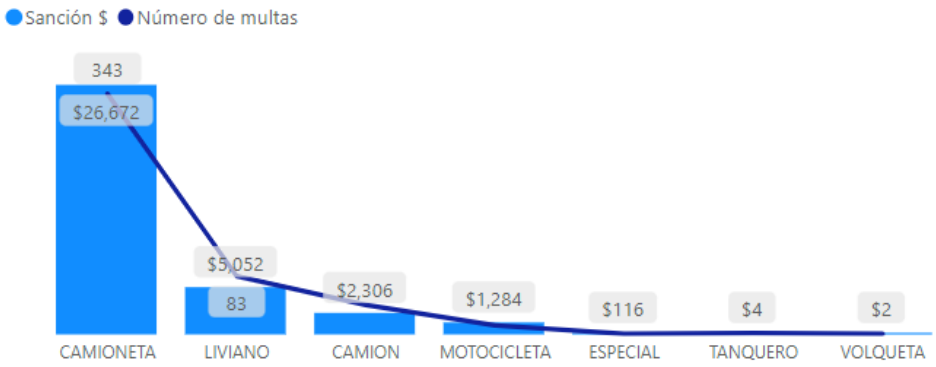


Figura 55 Número de infracciones y multa económica por clase de vehículos

La Gerencia de Operaciones al ser la que tiene asignada una mayor cantidad de camionetas, registra un mayor número de infracciones, de acuerdo a la Figura 56 concentra el 49% de las

mismas, seguida de esta se encuentra la Gerencia Comercial con el 14% y la Gerencia Técnica de Infraestructura con el 12%.

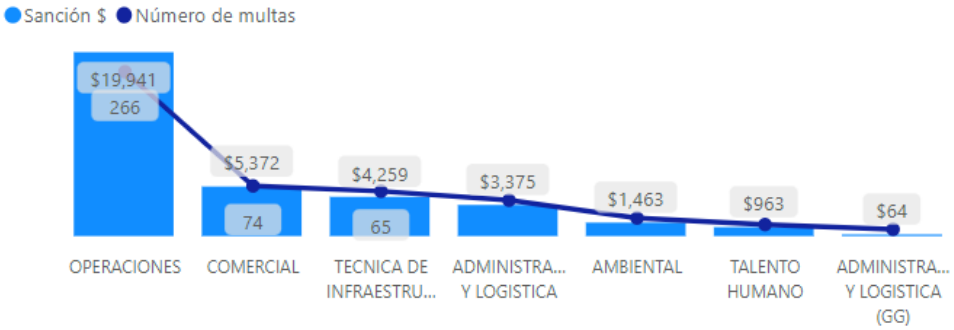


Figura 56 Número de infracciones y multa económica por Gerencia

En la Tabla 12 se identifican a los vehículos con mayor cantidad de infracciones, existen vehículos con hasta 15 infracciones.

Tabla 12 Top 10 - Ranking de vehículos infractores y multa económica

Placa	Articulo	Multas	Sancion	Clase
PME0033	EXCEDO VELOCIDAD - RANGO MODERADO	15	$1,774	CAMIONETA
PME0348	EXCEDO VELOCIDAD - RANGO MODERADO	10	$1,124	CAMIONETA
PMF0704	EXCEDO VELOCIDAD - RANGO MODERADO	8	$931	CAMIONETA
PCY4898	EXCEDO VELOCIDAD - RANGO MODERADO	7	$834	
PME0282	EXCEDO VELOCIDAD - RANGO MODERADO	7	$816	CAMIONETA
PDL9345	EXCEDO VELOCIDAD - RANGO MODERADO	7	$780	
PMD0995	EXCEDO VELOCIDAD - RANGO MODERADO	7	$774	CAMIONETA
PME0328	EXCEDO VELOCIDAD - RANGO MODERADO	5	$589	CAMIONETA
PMD0858	EXCEDO VELOCIDAD - RANGO MODERADO	5	$586	CAMIONETA
PME0258	EXCEDO VELOCIDAD - RANGO MODERADO	5	$580	CAMIONETA
Total		**545**	**$41,537**	

3.5. Modelo de clasificación de conductores

Los resultados del modelo generado se presentan en la Tabla 13, dicho modelo genera resultados similares a los modelos generados con *K-means*, esto es, cuatro grupos importantes y dos con menor cantidad de conductores, en donde sobresalen los grupos 1, 3 y 5 con mayor puntaje promedio, por el contrario, el grupo 4 con menor puntaje promedio.

El grupo 3 además registra la menor cantidad de puntos promedio perdidos, mientras que el grupo 4 presenta la mayor cantidad de puntos promedio perdidos.

Existe una mayor cantidad de puntos perdidos y sanción en vehículos particulares en los grupos 1, 2 y 5.

Los grupos 2, 3 y 1 registran un mayor monto en sanciones y sanciones promedio en vehículos de la flota.

El grupo 2 presenta un mayor número de citaciones promedio por conductor en los vehículos de la flota, mientras que para vehículos particulares lo hace el grupo 4.

Tabla 13 Resultados de modelo SVM balanceado 70%-30% en variables de la flota y de conductores

cluster_svm	N. cond.	Citaciones flota	Citaciones promedio flota	Puntos perdidos flota	Sancion	Promedio de Sancion flota	Promedio de puntos	Distancia promedio (km)	Tiempo promedio mov
1	84	48	0,6	0	$3.829	$80	27,45	15,30	4,06
2	107	257	2,4	6	$18.947	$74	25,64	21,71	5,66
3	55	69	1,3	0	$5.652	$82	27,75	24,51	9,00
4	8	10	1,3	6	$259	$26	11,63	18,98	6,00
5	77	44	0,6	6	$2.966	$67	28,62	18,50	8,69
6	1	1	1,0	9	$183	$183	24,00		
Total	**332**	**429**	**1,3**	**27**	**$31.836**	**$74**	**26,80**	**23,74**	**8,63**

cluster_svm	N. cond.	N. citaciones	Citaciones promedio conductor	Puntos perdidos	Puntos perdidos promedio conductor	Sancion	Promedio de Sancion	Promedio de puntos
1	84	118	1,40	139,50	1,66	$10.431	$88	27,45
2	107	116	1,08	108,00	1,01	$9.687	$84	25,64
3	55	65	1,18	52,50	0,95	$3.938	$61	27,75
4	8	25	3,13	49,50	6,19	$2.063	$83	11,63
5	77	82	1,06	82,00	1,06	$8.728	$106	28,62
6	1							24,00
Total	**332**	**406**	**1,22**	**431,50**	**1,30**	**$34.847**	**$86**	**26,80**

El modelamiento permitió identificar a los conductores más riesgosos, siendo éstos los conductores del grupo 2 y 4, quienes registran una mayor cantidad de citaciones promedio en la flota y en vehículos particulares respectivamente. El grupo 4 además, presenta una mayor cantidad de puntos perdidos promedio por conductor.

4. CONCLUSIONES Y RECOMENDACIONES

4.1. CONCLUSIONES

Los procesos de *web scraping* son clave en este estudio, ya que permitieron obtener información valiosa a un bajo costo computacional y en un tiempo reducido, dichos procesos a su vez permitieron reemplazar una tarea manual para un funcionario de la Empresa, la cual le tomaba alrededor de 15 días y dejaba un espacio para el error humano.

La metodología CRISP-DM, con la cual se desarrolló este estudio permitió crear un flujo claro para una mayor facilidad de comprensión e implementación, pasando por sus seis etapas, permitiendo conocer las necesidades del negocio, los datos a utilizar, la preparación requerida sobre éstos, para su posterior modelamiento y evaluación.

Las técnicas de aprendizaje automático no supervisado *K-means* y agrupamiento jerárquico, y no supervisado árbol de decisión y máquina de soporte vectorial empleadas permitieron clasificar a los conductores en base a sus características y comportamiento durante la conducción, el balanceo de datos aplicado, así como los distintos cortes en los grupos de entrenamiento y de prueba permitieron generar varios escenarios de análisis para evitar un sobreajuste de los modelos.

En relación a la evaluación de los modelos generados, al tratarse de un problema de clasificación se utilizaron matrices de confusión; y, varias métricas como la exactitud, la precisión, sensibilidad y *F1-score* por clase, los resultados de esta evaluación posibilitaron la selección de manera sustentada del mejor modelo.

El tablero desarrollado en este estudio permitió consolidar el análisis realizado de conductores, flota vehicular, infracciones a vehículos de la flota y particulares, utilización de la flota y el mapa de recorridos frecuentes, e integrarlo con el modelo de clasificación de conductores, al ser una herramienta de fácil manejo permite además realizar mayores análisis sobre la flota y conductores.

Dentro de las herramientas utilizadas en este estudio toma relevancia el lenguaje de programación R, ya que el mismo se utilizó durante todo el proceso, partiendo del *web scraping* en donde se integró con Python, pasando por los procesos de limpieza y minería de datos, el modelamiento y su correspondiente evaluación, finalizando con la integración con Power BI de un objeto visual desarrollado con dicho lenguaje, mismo que era imposible de realizar en Power BI por su limitado nivel de personalización.

A nivel del negocio se generan las siguientes conclusiones sobre las cuales se debería trabajar para lograr mejores resultados en la Empresa:

- A partir del modelo de clasificación de conductores se generan dos grupos de riesgo alto (grupos 2 y 4), con un total de 115 conductores sobre los cuales es necesario aplicar medidas correctivas para mitigar el nivel de riesgo durante la conducción, por el contrario, el grupo de mejores conductores (grupo 3) podría ser considerado para algún incentivo.
- El análisis de los conductores permitió identificar a 7 conductores que registran menos de 10 puntos en su licencia de conducción, por cuanto se recomienda tomar medidas como cursos de recuperación de puntos o cambios en el personal.
- El análisis de monitoreo satelital por su parte permitió identificar a los conductores que registran velocidades de conducción excesivas (dentro y fuera del rango moderado), sobre quienes es necesario aplicar acciones para cambiar este comportamiento.
- El análisis de la flota permitió identificar a los vehículos con mayor número de infracciones, los cuales en su mayoría son camionetas, así también se identificó que la mayor causal de infracciones es el exceso de velocidad, seguido de desobediencia de órdenes a los agentes de tránsito, las cuales están generando un valor importante en sanciones económicas para la Empresa.
- El análisis de tiempo de uso y distancia recorrida promedio por vehículo permitió identificar departamentos y Unidades con subutilización de sus vehículos, sobre los

cuales se recomienda analizar a mayor profundidad la necesidad de los mismos o alternativas para optimizar el uso de la flota en base a la realidad institucional.

- El mapa desarrollado para identificar las principales rutas de los vehículos por Gerencia permitió además identificar los principales destinos de los vehículos por clase, lo cual puede servir a su vez para una reasignación de estacionamientos en base a la naturaleza de los vehículos y su destino más frecuente.

4.2. RECOMENDACIONES

La extracción de datos mediante *webscraping* para este estudio fue esencial, las herramientas utilizadas para este fin fueron las adecuadas, ya que permitieron interactuar con las páginas de consulta y extraer la información de manera automática; sin embargo, se recomienda utilizar el lenguaje de programación Python para el desarrollo de procesos de *webscraping* frente a R, dado que existe mayor documentación y la comunidad de usuarios para este fin es superior, por otro lado, para la limpieza, análisis y modelamiento no existen diferencias relevantes ni limitantes.

La metodología *CRISP-DM* permitió generar un flujo de trabajo claro para la planificación y desarrollo de este estudio, por cuanto se recomienda su aplicación en temáticas similares.

Los algoritmos utilizados permitieron clasificar a los conductores en base a sus características y comportamiento, dicha clasificación en conjunto con el análisis descriptivo realizado se convierte en un instrumento para la toma de decisiones y la generación de estrategias de reducción de riesgos y optimización de la flota; sin embargo, se recomienda actualizar el modelo de clasificación luego de observar cambios importantes en el comportamiento de conductores o cuando exista una reasignación de vehículos, además se podrían añadir otros algoritmos para contrastar resultados.

Se recomienda a la Empresa objeto de este estudio capacitar al personal del área de Logística en las herramientas utilizadas a fin de dar continuidad a los análisis realizados y mantener procesos automáticos de consulta de información que permitan alcanzar una mayor eficiencia.

Es importante automatizar los procesos de extracción, transformación y carga de datos que alimentan el tablero desarrollado a fin de contar con información actualizada, de manera especial en el mapa, el cual a su vez podría migrar a una herramienta especializada en información geográfica como QGIS o ArcGIS para un mayor aprovechamiento de los datos del monitoreo satelital, permitiendo realizar análisis y síntesis de datos espaciales, así como profundizar en análisis muy específicos, identificar patrones y relaciones, entre otros.

5. REFERENCIA BIBLIOGRÁFICAS

[1] J. Gómez Berenguer, Desarrollo de modelo de analítica avanzada para optimización en transición de flota, Madrid: Universidad Pontificia Comillas, 2021.

[2] Y. Rodriguez y M. Pinto, Modelo de uso de información para la toma de decisiones estratégicas en organizaciones de información cubanas, Campinas: Transinformação, 2018.

[3] F. El Khoury y A. Zgheib, Building a Dedicated GSM GPS Module Tracking System for Fleet Management Hardware and Software, Boca Raton: Taylor y Francis Group, 2018.

[4] F. Pérez Maier, «RedGps,» 26 06 2020. [En línea]. Available: https://www.redgps.com/blog-noticias/evolucion-rastreo-gps-avl-al-iot. [Último acceso: 19 04 2023].

[5] Y. Yang, Z. Yuan, X. Fu, Y. Wang y D. Sun, Optimization Model of Taxi Fleet Size Based on GPS Tracking Data, Beijing: MDPI, 2019.

[6] J. Ascencio, A. Bustos, A. Zamora y J. Balbuena, Desarrollo de mapas resultado del asistente automático para el diseño de rutas de distribución, México: Instituto Mexicano del Transporte, 2022.

[7] J. L. de los Mozos Quiroga y S. Moreno López, Optimización de flotas de vehículos, una herramienta para incrementar la eficiencia, Madrid: Universia Business Review, núm. 16, 2007.

[8] Contraloría General del Estado, «Contraloría General del Estado,» 2 4 2003. [En línea]. Available: https://www.contraloria.gob.ec/WFDescarga.aspx?id=511&tipo=nor. [Último acceso: 19 4 2023].

[9] F. Nwanganga y M. Chapple, Practical Machine Learning in R, Indianapolis: Wiley, 2020.

[10] K. Sud, P. Erdogmus y S. Kadry, Introduction to Data Science and Machine Learning, Londres: IntechOpen, 2020.

[11] C. Prabhu, Fog Computing, Deep Learning and Big Data Analytics-Research Directions, Nueva Delhi: Springer, 2019.

[12] M. A. Khder, Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application, Bahrain: Int. J. Advance Soft Compu. Appl, Vol. 13,, 2021.

[13] R. Vording, Harvesting unstructured data in heterogenous business, Netherlands: University of Twente, 2021.

[14] K. Mehta, M. Salvi, R. Dand, V. Makharia y P. Natu, Comparative Study of Various Approaches to Adaptive, Springer, 2019.

[15] F. Mattosinho, Mining Product Opinions and Reviews on the Web, Dresden: Technische Universität Dresden, 2010.

[16] M. Gheorghe, F. Mihai y M. Dârdală, Modern techniques of web scraping for data scientists., International Journal of User-System Interaction, 2018.

[17] M. Ordoñez, Optimización de redes UMTS soportada en machine learning, Bogotá: Universidad Distrital Francisco José De Caldas, 2021.

[18] Sanchez, Automatización de diagnósticos mediante el uso de técnicas de Machine Learning, Sevilla: Universidad de Sevilla, 2021.

[19] C. Pérez, Prospección de fuga de clientes de un servicio de suscripción digital, Santiago: Universidad del Desarrollo, 2020.

[20] L. José, Clasificación de conductores a partir de la información de diagnóstico de un vehículo mediante técnicas de aprendizaje automático, Alicante: Universidad de Alicante, 2017.

[21] D. A. Silva Palacios, Clasificación Jerárquica Multiclase, Valencia: Universitat Politècnica de València, 2021.

[22] P. Chapman, J. Clinton, R. Keber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, CRISP-DM 1.0 Step by step guide, SPSS, 2000.

[23] T. Iglesias, Métodos de Bondad de Ajuste en Regresión Logística, Granada: Universidad de Granada, 2013.

[24] T. Hennig-Thurau y U. Hansen, Relationship Marketing: Gaining Competitive Advantage Trough Customer Satisfaction and Customer Retention, Hannover: Springer, 2020.

[25] J. L. De los Mozos Quiroga y S. Moreno López, Optimización de flotas de vehículos, una herramienta para incrementar la eficiencia, Madrid: Universia Business Review, núm. 16, 2007.

[26] S. Han, D. Williamson y Y. Fong, Improving random forest predictions in small datasets from two-phase sampling designs, BMC Medical Informatics and Decision Making, 2021.

[27] V. Krotov y M. F. Tennyson, Research Note: Scraping Financial Data from the Web Using the R Language, Journal of Emerging Technologies in Accounting , 2018.

[28] Red Hat, «Red Hat,» 12 11 2021. [En línea]. Available: https://www.redhat.com/es/topics/internet-of-things. [Último acceso: 19 4 2023].

[29] B. Golden, S. Raghavan y E. Wasil, The vehicle routing problem: latest advances and new challenges, New York: Springer, 2008.

[30] J. Velez, Diseño e implementación de técnicas metaheurísticas en transporte multimodal de mercancías con cadena de frio, Valencia: Universitat Politecnica de Valencia, 2022.

[31] A. Carrillo y G. Rocha, Diseño de un módulo de inteligencia artificial, para el reconocimiento facial de un chofer autorizado a conducir las unidades de transporte de la compañía transportes transplaneta s.a., evaluado en un ambiente de prueba, Quito: Universidad Politécnica Salesiana, 2022.

[32] P. López, Aplicacion de técnicas de inteligencia artificial para la prediccion de congestiones a corto plazo, Bilbao: Universidad de Deusto, 2016.

[33] A. Rosado y A. Verjel, Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander, Santander: Tecnura, 2015.