CLASIFICACIÓN DE LOS MUNICIPIOS DEL PAÍS DE ACUERDO CON EL MOVIMIENTO DE CARGA POR CARRETERA UTILIZANDO MODELOS NO SUPERVISADOS DE APRENDIZAJE AUTOMÁTICO

Juan Carlos Martinez Rodriguez, jcmartinezr01@libertadores.edu.co

RESUMEN

El presente documento desarrolla un modelo de clasificación no supervisado para los municipios de Colombia en torno al movimiento de carga del servicio público de transporte de carga, a través, de la utilización de algoritmos de clasificación como los son los árboles de clasificación, k-medias, k-medioides y modelo de clasificación discreta. Estos algoritmos son utilizados en conjunto con los datos de movilización de carga del sistema del Registro nacional de Despachos de Carga – RNDC, del Ministerio de Transporte de Colombia, de modo tal que se realiza un proceso de recolección de datos, preparación de los mismos en torno al flujo de carga, desarrollo de los algoritmos mencionados y una evaluación de los diferentes modelos para seleccionar aquel modelo que en términos de homogeneidad intra grupal y heterogeneidad intergrupal sea el más adecuado. De acuerdo con lo mencionado, el modelo seleccionado es k-medioides, este modelo utiliza

la mínima distancia de la mediana de los miembros de cada grupo, generando tres grupos cada uno con sus características particulares.
Palabras clave: Aprendizaje no supervisado, Clusterización, RNDC, Municipios.

ABSTRACT

This document develops a no supervised clustering models applied to the colombia's cargo flow public service, using clustering models as decision trees, k-means, k-medioids and DBSCAN. This algoritms are used with the data of the National Registry of Cargo Dispatches (RNDC) of the Colombian Ministry of Transportation. The process starts with recollecting the data, preparing the data, develop the algoritms and evaluate them to select the model that best fits in intra-group homogeneity and intergroup heterogeneity. According to the mentioned, k-medioids is the selected model, this model uses the median minimum distance between members of each group, generating three groups, each one with its own features.
Keywords: Non supervised machine learning, Clustering, RNDC, Township.

INTRODUCCIÓN

El Ministerio de Transporte es "una entidad del orden Nacional, encargada de garantizar el desarrollo y mejoramiento del transporte, tránsito y su infraestructura, de manera integral, competitiva y segura, buscando incrementar la competitividad del país, con tecnología y recurso humano comprometido y motivado" (Ministerio de Transporte, 2021). Dentro del Ministerio de Transporte, se encuentra el Grupo de Logística, cuyo objetivo es buscar la reducción de costos, tiempos y el impacto ambiental de los procesos de la logística de carga a través de la línea de acción de corredores eficientes (Ministerio de Transporte, 2019). De esta forma, la toma de decisiones en el sector del servicio

público de transporte de carga por carretera, es materia de búsqueda de diferentes aproximaciones para una gestión eficiente en la generación de política pública para el sector. El Ministerio de Transporte, al momento de definir políticas públicas para el sector de transporte de carga terrestre, requiere identificar los municipios cuyo nivel de impacto en estas decisiones se ven alta, mediana o mínimamente afectados. Hoy en día, la Decreto 2106 de 2019 en su artículo 153, clasifica los municipios por criterios de población e ingresos corrientes de libre destinación y esta clasificación no refleja la dinámica del flujo de carga del país. Para conocer esta distribución se requiere clasificar los municipios a partir de diferentes aspectos que involucran el flujo de carga para cada municipio, lo que se ve reflejado en los kilogramos, galones, viajes realizados y conexiones hechas.

Estas decisiones, al ser aplicadas de manera uniforme a todos los municipios del país, generan impactos adversos en municipios donde el volumen de carga es reducido respecto a ciudades principales; y, en las ciudades principales una decisión de este tipo, no generaría el efecto deseado dado que la rigurosidad y la exigencia de las mismas son reducidas respecto a las externalidades propias de cada ciudad. De acuerdo con lo anterior, es necesario determinar un modelo de clasificación de los municipios que se ajuste a variables como lo son kilogramos, viajes, galones o conexiones, las cuales están directamente relacionadas con el flujo de carga del sector.

Teniendo en cuenta esta necesidad y con la información pública que tiene el sistema del Registro Nacional de Despachos de Carga – RNDC –, se identifica la oportunidad de desarrollar un modelo de clasificación de los municipios del país basado en aprendizaje

automático no supervisado, teniendo en cuenta variables como kilogramos, galones, viajes y conexiones de cada uno de los municipios.

Esta clasificación de municipios por importancia de movimiento de carga, permite al Ministerio de Transporte generar políticas en el sector a la medida de cada municipio y gestionar necesidades conexas a este. Adicionalmente, permite la planeación organizada de modo que las características del movimiento de carga sean eficientes respecto a la clasificación de cada municipio en el país. Por esta razón, se propone desarrollar un modelo de clasificación de municipios de acuerdo con el movimiento de carga utilizando modelos de aprendizaje automático no supervisados de modo tal que encontremos municipios con características similares respecto a las variables mencionadas.

Para esta clasificación se utilizaron modelos no supervisados de aprendizaje automático y medidas estadísticas para determinar el número óptimo de grupos clasificados en cuanto a municipios resultados que se utilizará con herramientas georreferenciadas y modelos de red.

1. MARCO CONCEPTUAL

En esta sección se describen los conceptos necesarios para el desarrollo del modelo propuesto, de esta forma, se tiene la división territorial del país, el servicio público de transporte de carga por carretera, el aprendizaje automático, el agrupamiento y sus métodos, así como, las diferentes herramientas estadísticas para la determinación del número de grupos necesarios para el modelo, de forma tal que estos conceptos se enmarquen en la metodología propuesta para el desarrollo de este.

1.1. División territorial

Un municipio es la “entidad fundamental de la división político – administrativa del Estado” (República de Colombia, 1990), luego, a un nivel superior de los municipios, se encuentran los departamentos, que son entidades territoriales que “gozan de autonomía para la administración de los asuntos seccionales y la planificación y promoción del desarrollo económico y social dentro de su territorio en términos establecidos por la constitución. […] Los departamentos ejercen funciones administrativas, de coordinación, de complementariedad de la acción municipal, de intermediación entre la Nación y los Municipios (República de Colombia, 1990).

En Colombia hasta 2019, existen 1101 municipios organizados en siete categorías, de estas siete categorías, aproximadamente 957 municipios o el 86,9% son de la categoría sexta. Estas categorías se encuentran definidas en la ley 2106 de 2019 con tres tipos de municipios y siete categorías, determinadas a través de la población y los ingresos corrientes de libre destinación en la tabla 1 (República de Colombia, 2019).

De esta forma, la Nación al momento de formular política pública, se ve en la necesidad de establecer un diálogo con estas dos entidades, por una parte, con el departamento para lograr consenso y apoyo en la región que cubre el mismo y por otra parte, la acción directa sobre las disposiciones que la Nación defina que son aplicables para todos los municipios del país, como lo es la regulación y ámbito de aplicación del servicio público de transporte terrestre automotor de carga “considerado [como] un servicio público esencial” (Ministerio de Transporte, 1996) bajo regulación del Estado a través del Ministerio de Transporte.

Tabla 1. Categorización municipal del Decreto – Ley 2106 de 2019.

Tipo	Categoría	Población (habitantes)	Ingresos corrientes de libre destinación	Cantidad de municipios
Grandes municipios	Especial	Superior a 500.001	400.000 SMLMV	6
	1	Entre 100.001 y 500.000	Entre 100.000 y 400.000 SMLMV	25
Municipios intermedios	2	Entre 50.001 y 100.000	Entre 50.000 y 100.000 SMLMV	22
	3	Entre 30.0001 y 50.000	Entre 30.000 y 50.000 SMLMV	19
	4	Entre 20.0001 y 30.000	Entre 25.000 y 30.000 SMLMV	23
Municipios básicos	5	Entre 10.0001 y 20.000	Entre 15.000 y 25.000 SMLMV	49
	6	Igual o inferior a 10.000	No superiores a 15.000 SMLMV	957

Nota: Elaboración propia con base en (Departamento administrativo de la función pública, 2019) y (Contaduría General de la Nación, 2021).

En la tabla anterior se observa la distribución de los municipios a través de los criterios mencionados de población e ingresos corrientes de libre destinación presentándose de forma marcada en la categoría 6.

1.2. Servicio público de carga terrestre por carretera

El “servicio público de transporte terrestre automotor de carga. Es aquel destinado a satisfacer las necesidades generales de movilización de cosas de un lugar a otro, en vehículos automotores de servicio público a cambio de una remuneración o precio, bajo la responsabilidad de una empresa de transporte legalmente constituida y debidamente habilitada en esta modalidad, excepto el servicio de transporte de que trata el Decreto

2044 del 30 de septiembre de 1988 (Ministerio de Transporte, 2015). Este servicio público se monitorea a través del Registro Nacional de Despachos de Carga – RNDC – el cual es “el conjunto de datos relacionados con los vehículos de transporte de carga, con fines estadísticos para determinar la oferta de transporte de carga a nivel nacional” (Ministerio de Transporte, 2015). Estos datos se registran a través del registro administrativo llamado Manifiesto de carga, el cual es “el documento que ampara el transporte de mercancías ante las distintas autoridades, por lo tanto, debe ser portado por el conductor del vehículo durante todo el recorrido. Se utilizará para llevar las estadísticas del transporte público de carga por carretera dentro del territorio nacional.” (Ministerio de Transporte, 2015).

1.3. Aprendizaje automático

El aprendizaje automático conocido como (*machine learning)* es una rama de la inteligencia artificial que es capaz de convertir una muestra de datos en un programa informático capaz de extraer inferencias de nuevos conjuntos de datos para los que no ha sido entrenados previamente (Gonzalez-Marcos & Alba-Elias, 2017). El aprendizaje automático tiene diferentes aplicaciones que se permiten integrar en diferentes ámbitos de la sociedad en general, dentro de estos campos, la integración de estas técnicas en el sector público cada vez toma más relevancia debido al crecimiento de la información de carácter de interés general que requiere de análisis para la toma de decisiones, formulación de política pública o medición del impacto para el análisis de escenarios, de forma que el reto “no es la creación de una política de IA, sino su adecuada implementación y la materialización de sus principales compromisos, […] es decir, evidenciar el despliegue de la Inteligencia Artificial en el país a partir del cumplimiento de estos instrumentos” (Guio Español, 2020), adicionalmente, los diferentes tipos de

modelos existentes son del tipo supervisado y no supervisado (Gonzalez-Marcos & Alba-Elias, 2017).

El aprendizaje supervisado es el proceso donde se puede monitorear el resultado del proceso de aprendizaje del modelo, y, de acuerdo con los resultados se ajustan los hiper parámetros del modelo. El aprendizaje no supervisado no tiene una salida definida y de esta forma lo que se busca es encontrar un patrón de comportamiento de los datos asociados para generar al final de proceso un agrupamiento de los datos (Pérez Borrero & Gegúndez Arias, 2021).

El aprendizaje no supervisado se puede utilizar para diferentes tipos de objetivos, dentro de estos objetivos se encuentran los métodos de agrupamiento (*clusterización*), los cuales buscan identificar grupos de datos con semejanzas. En otros casos, se busca reducir la dimensionalidad de los datos para lograr un conjunto más manejable de variables, y, por otra parte, se puede ver el aprendizaje no supervisado como una extensión del análisis exploratorio de datos (Bruce, Bruce, & Gedeck, 2020).

1.4. Agrupamiento

El agrupamiento conocido en inglés como *clustering,* es una de las aplicaciones de aprendizaje automático que busca dividir los datos en grupos diferentes, donde los registros en cada grupo son similares. El objetivo del agrupamiento es identificar grupos de datos, los cuales pueden ser usados directamente, analizados en mayor detalle o llevados a un modelo de regresión o modelo de clasificación (Bruce, Bruce, & Gedeck, 2020). Entre los modelos a usar en esta aplicación se tienen los arboles de clasificación, k-medias, k-medioides y el modelo de densidad discreto entre otros.

1.5. Métodos de agrupamiento

Para el agrupamiento se tienen diferentes técnicas que a través del aprendizaje automático logran definir un conjunto de elementos que comparten elementos en común. Estos modelos van desde el k-medias, árboles de clasificación, k-medioides o el modelo de densidad discreto (DBSCAN) (Bruce, Bruce, & Gedeck, 2020) entre otros. Dado que las particiones no son siempre las mismas, cada modelo busca a través de su propia metodología, desarrollar un conjunto de agrupaciones lo suficientemente clara para encontrar las similitudes entre los individuos de cada grupo, de modo tal que estos agrupamientos permitan generar las respectivas evaluaciones de conveniencia de acuerdo con el contexto de aplicación de los modelos desarrollados.

- K-medias: El modelo de k-medias es un modelo de agrupamiento que busca a través de la minimización de la varianza total del sistema de acuerdo con la función potencial $\sum_i \quad \sum_j \quad d(x_{ij}, c_{i)}{}^2$ generar el agrupamiento de acuerdo con un parámetro k que define la cantidad de grupos a generar (Bruce, Bruce, & Gedeck, 2020).
- Árboles de clasificación: los árboles de clasificación están comprendidos en una familia genérica de algoritmos de agrupamiento que se construyen a través de grupos anidados por unión o partición sucesiva de los datos. Estos grupos son representados con un árbol o dendograma. En la parte superior del dendograma hay un único grupo que contiene todas las muestras, las hojas son los grupos de tamaño 1, y, la partición es hecha de acuerdo con la definición de un corte a una altura definida (Igual & Seguí, 2017).

- K-medioides: En el algoritmo de k-medioides la distancia Manhattan es usada como función objetivo, de esta forma, la función distancia es definida como $Dist(\underline{X_j}, \underline{Y_j}) = \| X_i - Y_j \|$ siendo $\underline{Y_j}$ la mediana de los puntos de datos en medio de cada dimensión del cluster Cj. Esto es porque el punto que tiene el mínimo de la suma de las L_1distncias a un conjunto de puntos distribuidos en una línea es la mediana de este conjunto (Aggarwal, 2015).
- Modelo de densidad discreto: El algoritmo DBSCAN del inglés *Density-Based Spatial Clustering of Applications with Noise* o modelo de densidad discreto es un algoritmo de agrupación, no paramétrico, basado en densidad de agrupamiento definida, este algoritmo busca un punto con suficiente densidad y va construyendo un clúster añadiendo puntos cercanos relevantes al mismo, luego, salta a otro punto con alta densidad y construye otro clúster independiente (Aggarwal, 2015).

2. METODOLOGÍA

La investigación desarrollada, es una investigación del tipo analítico respecto a la cual el problema de estudio consiste en la descomposición del todo, para determinar las causas, naturaleza del objeto de estudio y los efectos de la aplicación técnica en un hecho particular. De esta forma es posible lograr resultados en términos de explicación, desarrollo de analogías, comprensión del comportamiento y establecer nuevas teorías. (Ortiz Uribe & Garcia Nieto, 2000).

El proceso desarrollado se resume en actividades de recolección de datos, preparación de datos, desarrollo del modelo mediante técnicas de aprendizaje automático y evaluación del modelo. En términos generales, la recolección de datos se refiere al proceso de recolección de datos vinculados con el problema a analizar, la preparación

de los datos se refiere al proceso de preparación de los datos de acuerdo con las necesidades técnicas y del problema a analizar, la tercera etapa del proyecto se refiere a la ejecución de los algoritmos definidos para el problema a resolver y finalmente se realiza una evaluación de los modelos obtenidos a través de la ejecución de dichos algoritmos. Por otra parte, se debe tener en cuenta que el proceso de aprendizaje automático no supervisado, es un proceso de análisis de datos de modo que el aprendizaje y el conocimiento aplicado producen generalizaciones de los datos analizados. El proceso de aprendizaje automático se divide en un conjunto de fases que son: recolección de datos, preparación de los datos, aprendizaje (modelado) y evaluación del modelo.

2.1. Recolección de datos

En esta fase, se requiere recolectar la mayor cantidad de datos relacionados con el problema a resolver, en esta fase se recolectaron los datos generados a partir de la base pública del sistema de Registro Nacional de Despachos de Carga – RNDC. Este sistema recibe la información de los viajes realizados por los vehículos de carga del servicio público de carga por carretera, aquí en esta fase se realizó la recopilación de datos del año 2020, se realizó la descarga de los datos de la base y se agruparon los datos mensualmente.

2.2. Preparación de datos

La segunda fase es la preparación de datos, consistió en adecuar los datos obtenidos en la primera fase, esta preparación permitió obtener nuevas características de los datos a partir de las transformaciones realizadas, de acuerdo con las características que cuenta la información procesada. Dentro de la transformación de los datos, se realiza la

agregación de los valores por nodos, aplicando sumas o conteos de los registros individuales, adicionalmente, se realiza un análisis descriptivo de las variables obtenidas, luego, una normalización de los datos, un análisis de correlación y el análisis estadístico para determinar la cantidad de grupos necesarios para proceder a alimentar los modelos definidos.

2.3. Desarrollo del modelo mediante técnicas de aprendizaje automático

Luego de preparar los datos, se realiza la ejecución de los algoritmos definidos para el problema a resolver, estos algoritmos se ejecutan de acuerdo con las necesidades de procesamiento de información, a partir de los resultados de los modelos se obtienen conclusiones, identificando patrones o tendencias presentes en los datos.

Se realiza la selección de cuatro modelos para el análisis, partiendo desde modelos ampliamente usados, estos modelos son el k-medias, k-medioides, árboles de clasificación y modelo de densidad discretos. El criterio de agrupación de k-medias es la mínima varianza entre los grupos formados, el criterio de k-medioides es la mediana de la distancia de los grupos respecto a un centroide y el árbol de clasificación es el mínimo de la distancia entre una función objetivo, en este caso el error de la suma de los cuadrados o varianza. de acuerdo con la utilización de estadísticos propios de cada algoritmo. El algoritmo de modelo de densidad discreto, o DBSCAN, utiliza un parámetro de radio de separación entre miembros y densidad mínima de grupo. Estos modelos entonces son desarrollados con una base de datos preparada para este ejercicio y los algoritmos son ejecutados en el software estadístico R en su versión 4.1.2 (bird hippie) para la plataforma Windows 10, x86, 64-bit. Finalmente se utilizan las librerías readxl,

MVN, dplyr, pcaPP, cluster, factoextra, FactoMineR, ggplot2, Rcpp, dbscan, clvalid, corrplot, RColorBrewer y datasets (R Core Team, 2021).

2.4. Evaluación del modelo

Finalmente, la cuarta y última etapa es necesario evaluarla desde el punto de vista del problema original para determinar si la solución encontrada es lo suficientemente buena como para ser implementada. Luego de desarrollar los modelos, se procederá a determinar si los grupos definidos son homogéneos al interior y heterogéneos entre ellos, lo cual se confirmará a través de pruebas de hipótesis de igualdad de medias entre los grupos identificados en cada modelo, y de esta forma determinar la relación de cada grupo resultante de la clasificación con el movimiento de carga de cada municipio.

3. RESULTADOS

Como se menciona al inicio del documento, se desarrolló un modelo no supervisado de clasificación para determinar importancia del movimiento de carga de los municipios del país, este modelo utiliza la base de datos pública del sistema Registro Nacional de Despachos de Carga – RNDC, se realiza una transformación a partir de la agregación de los datos, se normaliza la información y se desarrollan diferentes modelos de clasificación en el software R para generar un conjunto de municipios que comparten características similares en cuanto al flujo de carga.

3.1. Recopilación de datos

Se tomó de la base de datos del RNDC, la información proveniente de los registros del año 2020 equivalente a 8’165.337 registros, de los viajes realizados por los vehículos, distribuidos en 15 variables descriptivas. A continuación, se presenta el esquema de los encabezados de la base de datos original.

Tabla 2. Diccionario de datos de la base del sistema RNDC.

Nombre variable	Descripción	Tipo
INGRESOID	Identificación única del registro individual, cada registro corresponde a un viaje realizado.	Numérico
MES	Año/mes de fecha de expedición del manifiesto	Numérico
COD_CONFIG_VEHICULO	Código configuración del vehículo de acuerdo con el número de ejes del cabezote y remolque o semiremolque, resolución 4100 2004. Código de la Configuración Resultante (Vehículo + Remolque)	Texto
CONFIG_VEHICULO	Descripción del código de la configuración del vehículo.	Texto
CODOPERACIONTRANSPORTE	Código del tipo de operación que se realiza para el transporte de mercancías.	Texto
OPERACIONTRANSPORTE	Descripción del Código del tipo de operación que se realiza para el transporte de mercancías	Texto
CODTIPOCONTENEDOR	Código que identifica si el contenedor viaja vacío o con carga	Texto
TIPOCONTENEDOR	Describe si el contenedor viaja vacío o con carga	Texto
CODMUNICIPIOORIGEN	Código DIVIPOLA del municipio donde se origina la operación	Numérico
CODMUNICIPIODESTINO	Código DIVIPOLA del municipio destino de la operación	Numérico
CODMERCANCIA	Código de la mercancía transportada de acuerdo a sistema armonizado. Código de identificación de la Carga. Relacionado con la naturaleza y características de la Carga.	Texto
KILOGRAMOS	Kilogramos transportados por cada viaje realizado	Numérico
GALONES	Galones transportados por cada viaje realizado	Numérico
KILOMETROS	Kilómetros entre el codmunicipoorigen y el codmunicipiodestino	Numérico
MANCANTIDADREMESAS	Cantidad de remesas relacionadas con el manifiesto de carga	Numérico

Estos campos describen la operación de transporte realizada por un vehículo de transporte de carga, de acuerdo con una selección basada en las características de la teoría de grafos donde se definen nodos, y aristas.

Los nodos en este caso son los municipios del país, los cuales pueden atraer o generar carga, bien sea en términos de galones o kilogramos y en base en este flujo, se tiene un balance de carga del municipio. Cada nodo se conecta por los viajes realizados entre municipios y se conecta con todos los municipios que haya realizado por lo menos 1 viaje, de esta forma se realiza la agregación de las variables teniendo en cuenta los criterios seleccionados.

3.2. Preparación de los datos

Se realizó la preparación de los datos a partir de las variables de viajes, galones, kilogramos y remesas, de esta forma se genera un conjunto de variables teniendo en cuenta que el municipio es un nodo dentro de una red de distribución del orden nacional para el servicio de carga.

Tabla 3. Diccionario de datos de la base de datos agregada.

Nombre variable	Descripción	Tipo
Mun_origen	campo DIVIPOLA de identificación de los municipios de origen considerados como nodos de entrada y salida de carga.	Numérico, identificador
Viajes_origen	Conteo de los viajes realizados con origen en el municipio.	Numérico
Kg_Muni_origen	Sumatoria de los kilogramos de los viajes realizados en el origen del municipio.	Numérico
Gal_Muni_origen	Sumatoria de los galones de los viajes realizados en el origen del municipio.	Numérico
Rem_Muni_origen	Sumatoria de las remesas de los viajes realizados en el origen del municipio.	Numérico
Nodos_origen	Conteo de la cantidad de destinos de los viajes originados desde el municipio	Numérico
Viajes_destino	Conteo de los viajes realizados con destino al municipio.	Numérico
Kg_Muni_destino	Sumatoria de los kilogramos de los viajes realizados hacia el municipio.	Numérico
Gal_Muni_destino	Sumatoria de los galones de los viajes realizados hacia el municipio.	Numérico
Rem_Muni_destino	Sumatoria de las remesas de los viajes realizados hacia el municipio.	Numérico
Nodos_destino	Conteo de la cantidad de orígenes que tienen viajes hacia el municipio.	Numérico
Municipio	Nombre del municipio asociado al DIVIPOLA	Texto
Departamento	Nombre del departamento asociado al DIVIPOLA	Texto

La tabla creada describe los movimientos de carga en cada municipio, de forma general se presentan 1102 registros, 13 variables, los estadísticos descriptivos se presentan en la tabla a continuación.

Tabla 4. Estadísticas descriptivas de la base de datos agregada. Parte 1.

	Viajes_origen	Kg_Muni_origen	Gal_Muni_origen	Rem_Muni_origen	Nodos_origen
Mínimo	0	0.000e+00	0	0.0	0.00
Cuartil 1	10	7.073e+04	0	9.0	5.00
Mediana	76	7.227e+05	0	76.0	16.00
Media	7410	9.864e+07	2864853	9296.2	54.31
Cuartil 3	891	1.038e+07	4300	884.8	47.00
Máximo	786526	1.173e+10	232866074	1201213.0	1010.00

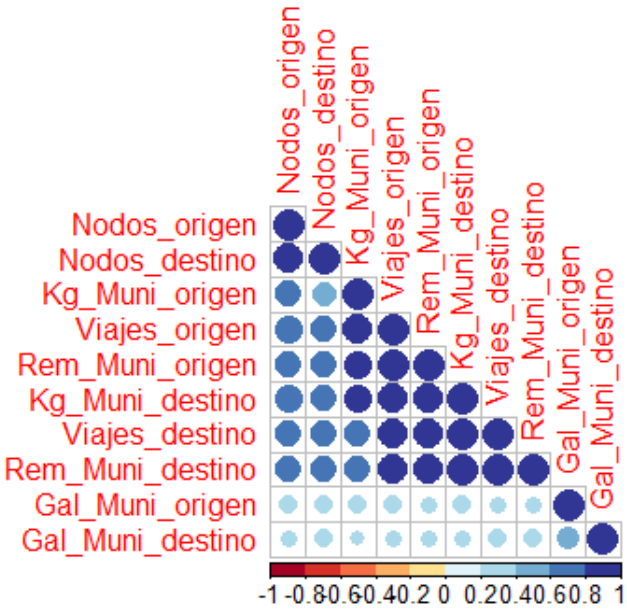
Se observa que los valores de la variable Gal_Muni_origen se encuentran muy agrupados, la mediana se encuentra en 0 y por lo tanto se puede intuir que la carga liquida no se origina en todos los municipios con registro.

Tabla 5. Estadísticas descriptivas de la base de datos agregada. Parte 2.

	Viajes_destino	Kg_Muni_destino	Gal_Muni_destino	Rem_Muni_destino	Nodos_destino
Mínimo	0.0	0.000e+00	0	0.0	0.00
Cuartil 1	144.0	9.951e+05	0	186.0	14.00
Mediana	540.5	4.434e+06	12000	713.5	32.00
Media	7409.6	9.864e+07	2864853	9296.2	54.31
Cuartil 3	2371.8	2.511e+07	293420	2922.8	68.00
Máximo	917300.0	1.057e+10	313870066	1183436.0	719.00

Estos datos presentados, se normalizan y se realiza la prueba de correlación para encontrar la medida de asociación entre las variables. De esta forma se presenta que la correlación es alta entre las variables de kilogramos, y viajes, bien sea en sentido de origen o sentido de destino.

Figura 1. Mapa de correlación de variables de la base normalizada.



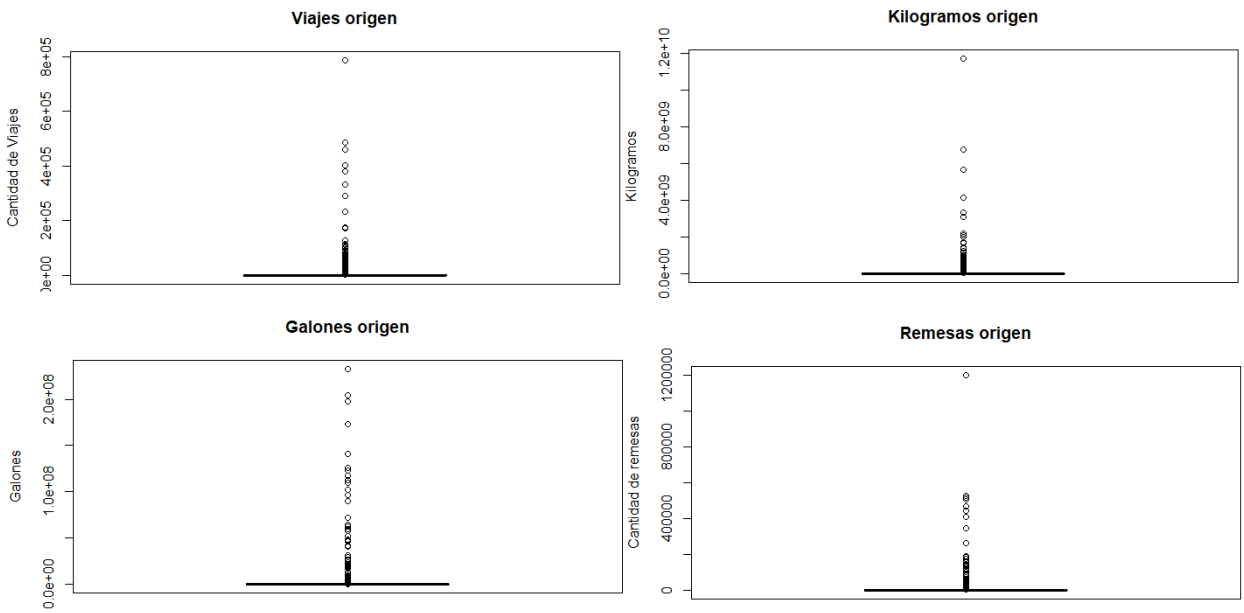
Para aclarar y ampliar la observación se presenta la matriz de correlación de la base de datos normalizada.

Tabla 6. Matriz de correlación de las variables de la base de datos normalizada.

	Viajes_origen	Kg_Muni_origen	Gal_Muni_origen	Rem_Muni_origen	Nodos_origen	Viajes_destino	Kg_Muni_destino	Gal_Muni_destino	Rem_Muni_destino	Nodos_destino
Viajes_origen	1,0000	0,8848	0,3378	0,9892	0,7462	0,9230	0,9548	0,2978	0,9009	0,6886
Kg_Muni_origen	0,8848	1,0000	0,3209	0,8213	0,6586	0,7304	0,8155	0,2412	0,6952	0,5886
Gal_Muni_origen	0,3378	0,3209	1,0000	0,3042	0,3468	0,3129	0,3284	0,4939	0,2977	0,3742
Rem_Muni_origen	0,9892	0,8213	0,3042	1,0000	0,7293	0,9432	0,9538	0,2861	0,9268	0,6788
Nodos_origen	0,7462	0,6586	0,3468	0,7293	1,0000	0,7171	0,7458	0,2996	0,7052	0,9150
Viajes_destino	0,9230	0,7304	0,3129	0,9432	0,7171	1,0000	0,9786	0,3423	0,9972	0,7253
Kg_Muni_destino	0,9548	0,8155	0,3284	0,9538	0,7458	0,9786	1,0000	0,3132	0,9669	0,7324
Gal_Muni_destino	0,2978	0,2412	0,4939	0,2861	0,2996	0,3423	0,3132	1,0000	0,3261	0,3325
Rem_Muni_destino	0,9009	0,6952	0,2977	0,9268	0,7052	0,9972	0,9669	0,3261	1,0000	0,7194
Nodos_destino	0,6886	0,5886	0,3742	0,6788	0,9150	0,7253	0,7324	0,3325	0,7194	1,0000

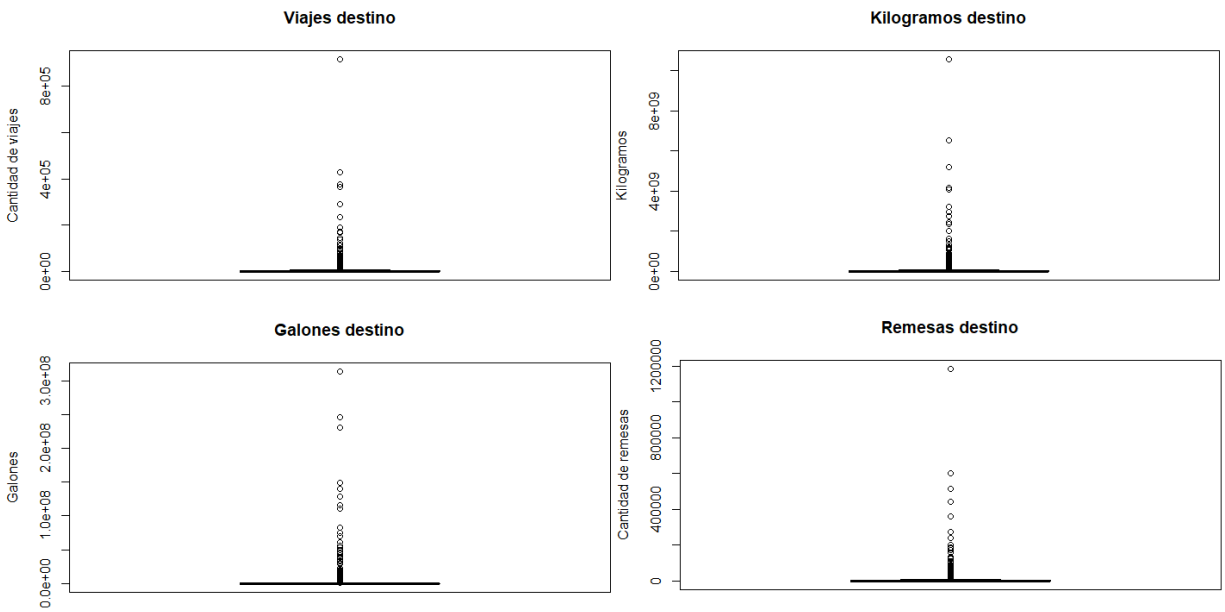
Cabe observar que la correlación de las variables Gal_Muni_origen y Gal_Muni_destino con las otras variables de análisis son valores que no superan el valor de 0.5 de modo tal que la correlación de las variables no es fuerte con las demás variables de análisis. Para identificar el comportamiento de las variables se realiza un análisis boxplot de las variables originales.

Figura 2. Boxplot de las variables de origen.



Se observa desde la gráfica, la presencia de valores atípicos, estos valores debido a la naturaleza de la información representan a los municipios que mayores valores presentan con el movimiento de la carga, es decir, son ciudades capitales, portuarias o centros industriales.

Figura 3. Boxplot de las variables de destino



También se observa que los valores 0 en las variables de galones, representan a los municipios donde no existe un origen o destino de movimiento de mercancía del tipo líquida, de modo tal que el movimiento por ejemplo de hidrocarburos no se presenta allí.

La presencia de datos atípicos genera sensibilidad en el resultado de la aplicación de la aplicación de cualquier metodología estadística que no sea robusta. Para el desarrollo del modelo se mantienen los datos atípicos, dado que corresponden al flujo de carga de cada municipio.

Los valores atípicos superiores de las variables, representan municipios cuya importancia en el flujo de carga en el país es superior, estos municipios pueden ser del tipo urbe con alta población, ciudad portuaria o centro industrial, y de esta forma es necesario mantenerlos dentro del desarrollo de los modelos.

Finalmente, se realizan el test de normalidad univariada para determinar la normalidad en alguna de las variables de la tabla normalizada, como resultado de las pruebas, se presentan los siguientes resultados.

Tabla 7. Test Kolmogorov – Smirnoff con variación de Lilliefors para prueba de normalidad univariada.

	Test	Variable	Statistic	P	value	Normality
1	Lilliefors	(Kolmogorov-Smirnov)	Viajes_origen	0.4285	<0.001	NO
2	Lilliefors	(Kolmogorov-Smirnov)	Kg_Muni_origen	0.4301	<0.001	NO
3	Lilliefors	(Kolmogorov-Smirnov)	Gal_Muni_origen	0.4580	<0.001	NO
4	Lilliefors	(Kolmogorov-Smirnov)	Rem_Muni_origen	0.4334	<0.001	NO
5	Lilliefors	(Kolmogorov-Smirnov)	Nodos_origen	0.3056	<0.001	NO
6	Lilliefors	(Kolmogorov-Smirnov)	Viajes_destino	0.4254	<0.001	NO
7	Lilliefors	(Kolmogorov-Smirnov)	Kg_Muni_destino	0.4247	<0.001	NO
8	Lilliefors	(Kolmogorov-Smirnov)	Gal_Muni_destino	0.4359	<0.001	NO
9	Lilliefors	(Kolmogorov-Smirnov)	Rem_Muni_destino	0.4272	<0.001	NO
10	Lilliefors	(Kolmogorov-Smirnov)	Nodos_destino	0.2146	<0.001	NO

De acuerdo con estos resultados, se encuentra que ninguna de las variables de la tabla normalizada tiene un comportamiento normal, y también es posible inferir la presencia de polos atractores o generadores de carga dentro del comportamiento del movimiento de carga en el país.

3.3. Desarrollo del modelo mediante técnicas de aprendizaje automático

Como tercera etapa del proceso, se debe determinar el número óptimo de grupos a obtener con los algoritmos de agrupación definidos para el ejercicio, este conjunto de grupos, se estiman mediante tres métodos definidos, los cuales son índice de silueta, mínimo de residuos cuadrados internos o WSS y el estadístico GAP.

Los algoritmos definidos para el desarrollo del modelo, se tienen los algoritmos de árboles de clasificación, k-medias, k-medioides y modelo de densidad discreto, a estos algoritmos, se les debe aplicar los métodos de definición del número óptimo de grupos para de esta manera realizar el ajuste de los parámetros o hiper parámetros, a excepción del modelo de densidad discreto.

A continuación, se presentan los métodos aplicados por cada algoritmo.

Figura 4. Método de índice de la silueta para algoritmo Ward.d2.

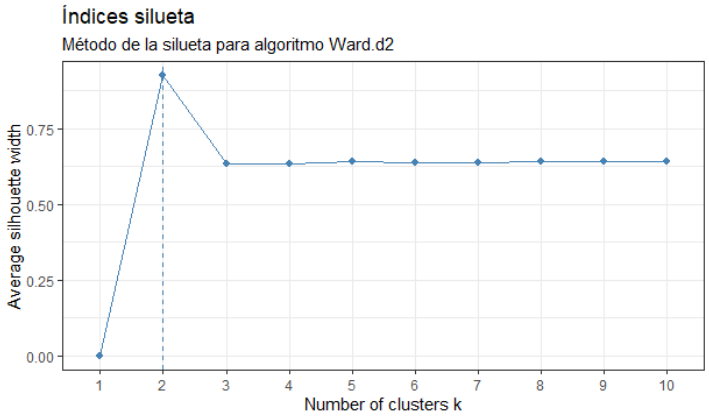


Figura 5. Método de índice de la silueta para algoritmo K-medias

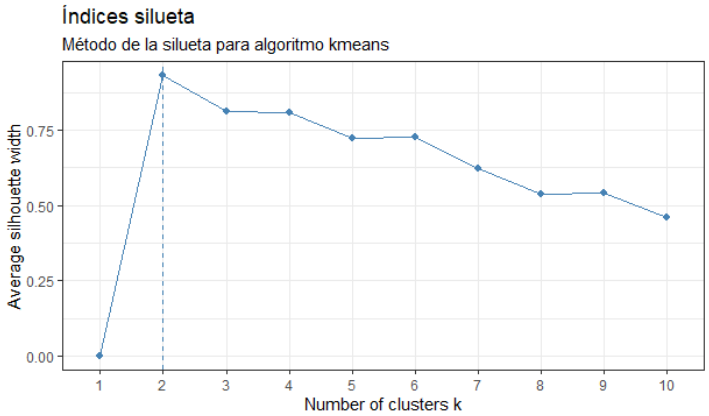
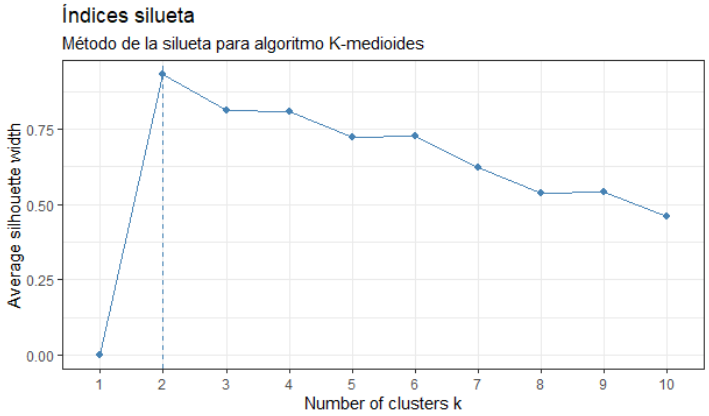


Figura 6. Método de índice de la silueta para algoritmo K-medioides.



Se observa que el número óptimo de agrupamientos que sugiere, es de 2 grupos.

Figura 7. Método WSS para algoritmo Ward.D2.

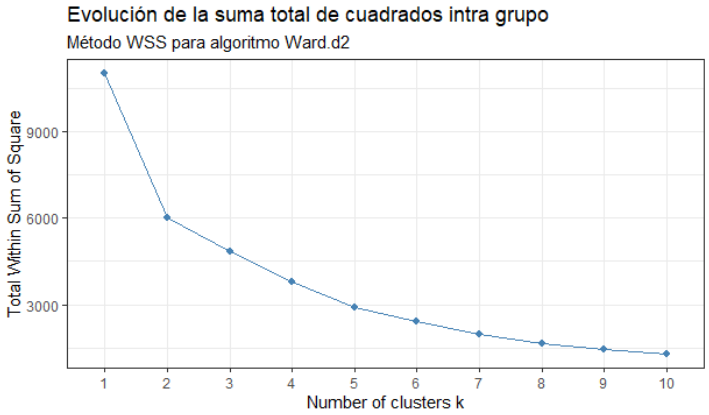


Figura 8. Método WSS para algoritmo k-medias.

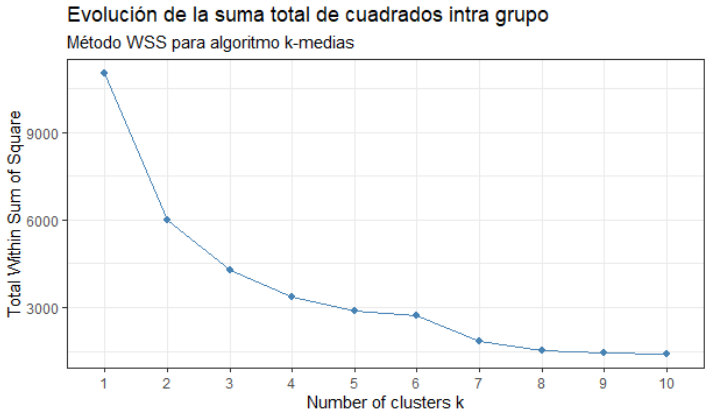
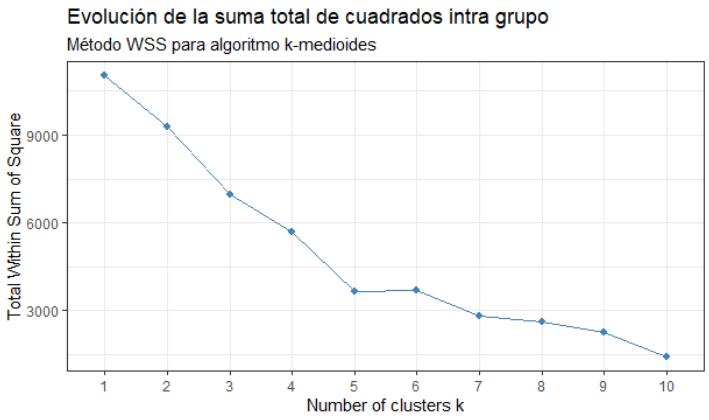


Figura 9. Método WSS para algoritmo k-medioides.



En este método se sugiere que sean 3 agrupamientos dado que en este punto la suma de los cuadrados intra-grupo no desciende más rápidamente.

Finalmente, se presentan los datos del método GAP para los tres algoritmos a los cuales se realiza el análisis.

Figura 10. Estadística GAP para el algoritmo Ward.D2.

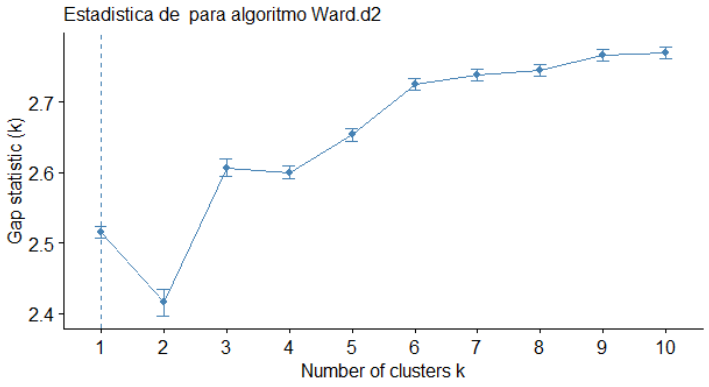


Figura 11. Estadística GAP para el algoritmo k-medias.

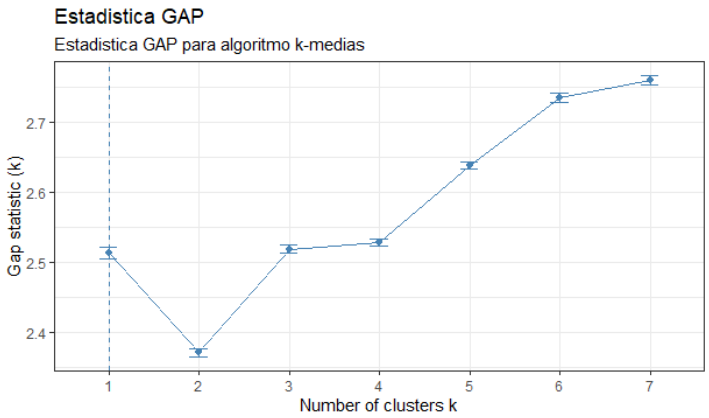
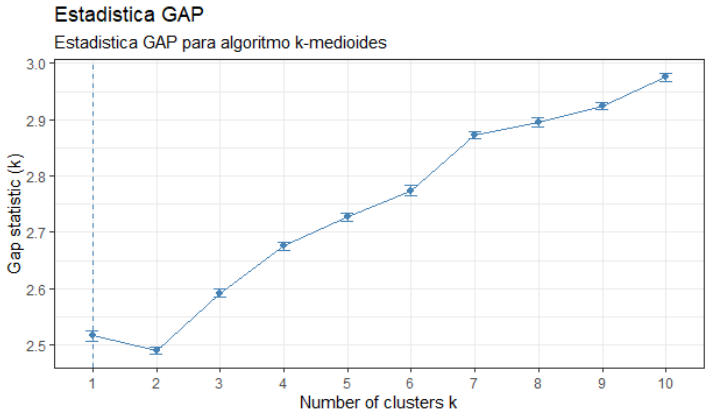


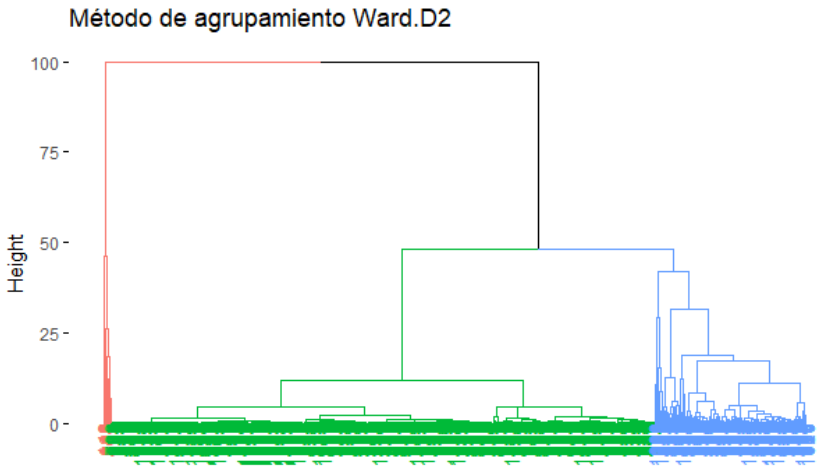
Figura 12. Estadística GAP para el algoritmo k-medioides.



Esta estadística considera como valor óptimo la creación de un solo grupo, pero si se analiza con detenimiento, la opción de 3 grupos es una opción viable para el desarrollo del ejercicio teniendo en cuenta las características del movimiento de carga en el país.

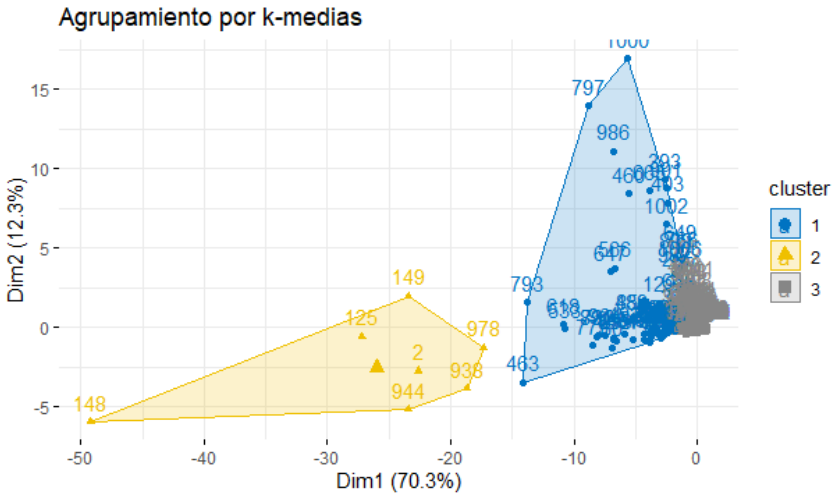
Se toma como opción el agrupamiento con 3 grupos de clasificación, entonces, se desarrollan los algoritmos de clasificación, inicialmente se ejecuta el algoritmo de árbol de clasificación Ward.D2, luego, el algoritmo k-medias, el algoritmo k-medioides y finalmente el algoritmo de modelo de densidad discreto.

Figura 13. Dendograma algoritmo de clasificación Ward.D2.



Se observa en el modelo del árbol de clasificación, los tres grupos definidos donde existe un grupo amplio, uno de tamaño medio y uno pequeño, estos datos serán analizados en conjunto con los resultados de los demás modelos.

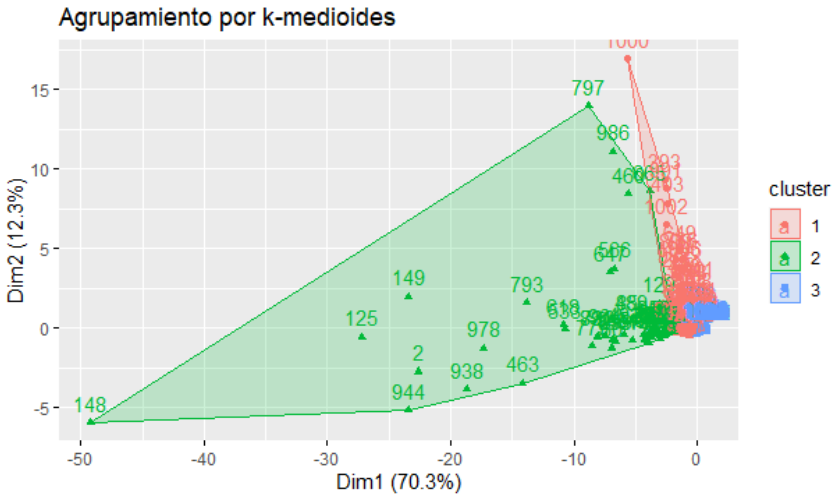
Figura 14. Agrupamiento por algoritmo de k-medias.



El modelo de agrupamiento del modelo de k-medias es un modelo que permite identificar claramente 3 grupos de datos dentro de la nube de datos representada gráficamente, de

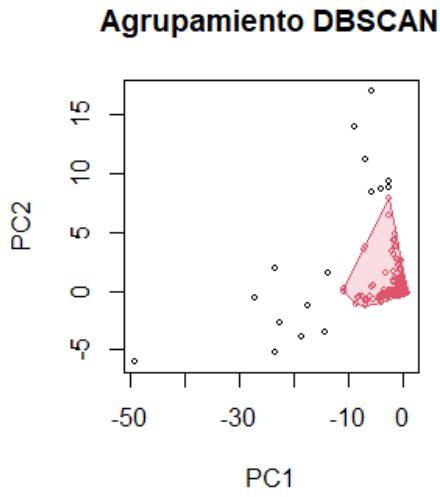
esta forma se evidencian los respectivos grupos de la clasificación. Existe en grupo 1 que es un grupo muy cercano al grupo 3, sin embargo, la densidad de agrupamiento del 3 es mucho más alta que el grupo 1, y de forma independiente se observa en la gráfica el grupo 2 identificado con el color amarillo.

Figura 15. Agrupamiento por k-medioides



Para el modelo de agrupamiento de k-medioides, los tres grupos generados se encuentran menos diferenciados, se observa que siguen el mismo patrón de agrupamiento del algoritmo de k-medias, aunque en este se observa más densidad que en el modelo anterior.

Figura 16. Agrupamiento por algoritmo de modelo de densidad discreto.



El modelo de agrupamiento de densidad discreto muestra únicamente un grupo creado y un conjunto de datos considerados ruido por el algoritmo, para efectos prácticos se analizará el ruido como un grupo y de modo tal que se contraste el resultado de la respectiva prueba. Se realiza el cálculo de los valores promedio de las variables de los modelos de agrupamiento y se presentan por cada modelo y por cada grupo generado.

Tabla 8. Valores promedio de las variables agrupadas en los modelos árboles de decisión y k-medias.

	Árbol de decisión			**K-medias**		
	Grupo 3	Grupo 1	Grupo 2	Grupo 3	Grupo 1	Grupo 2
Municipios de origen	855	236	11	1025	70	7
%	77,59%	21,42%	1,00%	93,01%	6,35%	0,64%
Promedio de Viajes_origen	372	16.948	349.812	1.273	54.573	434.340
Promedio de Kg_Muni_origen	8.701.363	219.018.550	4.506.975.284	22.598.808	659.226.637	5.627.875.591
Promedio de Gal_Muni_origen	54.564	10.853.794	49.901.865	790.015	28.318.367	52.145.305
Promedio de Rem_Muni_origen	375	20.824	455.373	1.336	70.000	567.909
Promedio de Nodos_origen	15	170	645	31	331	707
Promedio de Viajes_destino	694	17.642	309.854	1.855	49.777	397.020
Promedio de Kg_Muni_destino	6.785.039	231.347.176	4.391.420.897	21.701.267	694.419.675	5.407.370.829

Promedio de Gal_Muni_destino	141.320	11.083.494	38.230.393	750.010	28.849.536	52.691.465
Promedio de Rem_Muni_destino	884	21.998	390.652	2.342	61.642	504.105
Promedio de Nodos_destino	28	133	412	41	212	455

Para esta tabla se presentan los cálculos de los modelos del árbol de clasificación con el método Ward.D2 y el modelo k-medias. Se observa que el agrupamiento en el árbol de decisión se hace en 78-21-1 en términos de porcentaje de la distribución de los nodos analizados. Para el modelo de k-medias la distribución resulta en 93-6-1 de modo que el grupo mayor contiene aun más de 1000 nodos en un solo grupo.

Tabla 9. Valores promedio de las variables agrupadas en los modelos de k-medioides y DBSCAN.

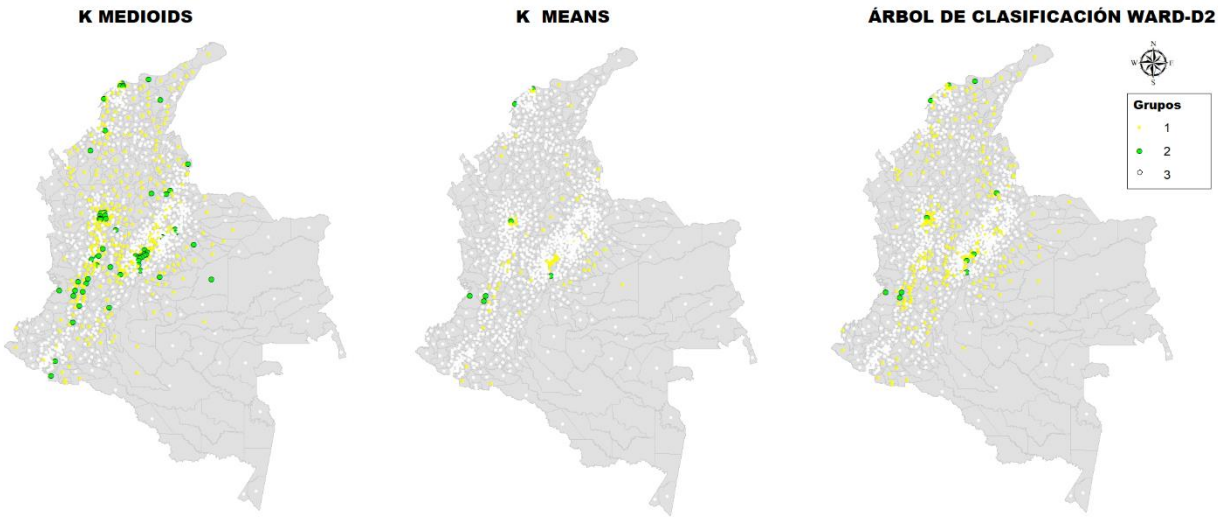
	K-medioides			**DBSCAN**	
	Grupo 3	Grupo 1	Grupo 2	Grupo 1	Ruido
Municipios de origen	710	331	61	1086	16
%	64,43%	30,04%	5,54%	98,55%	1,45%
Promedio de Viajes_origen	221	4.178	108.613	4.028	236.938
Promedio de Kg_Muni_origen	4.515.868	67.668.739	1.362.297.595	57.747.235	2.874.454.689
Promedio de Gal_Muni_origen	75.354	4.948.310	24.027.466	1.703.117	81.717.649
Promedio de Rem_Muni_origen	223	4.367	141.645	4.902	307.568
Promedio de Nodos_origen	10	82	422	48	483
Promedio de Viajes_destino	403	5.796	97.722	4.328	216.588
Promedio de Kg_Muni_destino	3.740.153	68.272.752	1.368.048.894	57.384.780	2.899.056.304
Promedio de Gal_Muni_destino	112.201	5.582.209	20.158.919	1.410.818	101.557.473
Promedio de Rem_Muni_destino	516	7.126	123.273	5.425	272.082
Promedio de Nodos_destino	21	89	260	50	320

Para el modelo de k-medioides se presenta una distribución de 64-30-6 en la distribución porcentual de los municipios, teniendo aquí la mayor asignación para el grupo 2 de los tres modelos, y finalmente, el modelo de densidad discreto que tiene una distribución de 99-1, lo que hace que sea un agrupamiento poco atractivo.

Para probar la conveniencia de los modelos, en la sección 3.4. se realizará la prueba de evaluación del comportamiento de las medias de las variables en los grupos de los modelos.

Para ilustrar la distribución de los municipios de acuerdo con el agrupamiento, se presenta la siguiente figura para notar la distribución espacial de los grupos.

Figura 17. Mapa de distribución de los nodos de acuerdo con el agrupamiento designado.



Se observa con cada algoritmo de agrupamiento, la localización de los nodos de importancia alta en color verde, los de media importancia en color amarillo y finalmente los nodos menos importantes en color blanco. La ubicación de los nodos de color verde se refiere a los centros poblados o zonas portuarias o centros industriales donde se

genera el principal movimiento de carga para el país y se encuentra entonces que los modelos siguen un patrón de clasificación y propagación de la importancia de cada nodo y que en unos casos se va ampliando con cada modelo ejecutado. Por otra parte, el modelo de densidad discreto considera los nodos importantes como ruido dentro del modelo, por lo tanto, no se consideró ser ilustrado para el documento.

3.4. Evaluación del modelo

Teniendo en cuenta los resultados de los modelos, se procederá a comparar los diferentes resultados con las diferencias de medias que se den en los diferentes grupos generados, de esta forma, la primera comparación que se realiza es con base en la lista de clasificación de municipios de la Ley 2106 de 2019 y seguidamente la comparación intra grupo. Los nodos del grupo 2 se enumeran en la siguiente tabla.

Tabla 10. Municipios clasificados en grupo 2 por modelo de clasificación.

#	K-medias	Decreto 2106 -2019	Resumen
1	MEDELLIN	Especial	
2	BARRANQUILLA	Especial	
3	BOGOTA, D.C.	Especial	
4	CARTAGENA DE INDIAS	Especial	
5	CALI	Especial	
6	BUENAVENTURA	Tipo 2	
7	YUMBO	Tipo 1	

Tipo de municipio	Cantidad	%
Especial	5	71%
Tipo 1	1	14%
Tipo 2	1	14%
Total general	7	100%

#	Árbol de clasificación Ward.D2	Decreto 2106 -2019	Resumen
1	MEDELLIN	Especial	
2	BARRANQUILLA	Especial	
3	BOGOTA, D.C.	Especial	
4	CARTAGENA DE INDIAS	Especial	
5	FUNZA	Tipo 1	
6	TOCANCIPA	Tipo 2	
7	SANTA MARTA	Tipo 1	
8	BUCARAMANGA	Especial	
9	CALI	Especial	
10	BUENAVENTURA	Tipo 2	

Tipo de municipio	Cantidad	%
Especial	6	54.55%
Tipo 1	3	27.27%
Tipo 2	2	18.18%
Total general	11	100.00%

11	YUMBO	Tipo 1	
#	**K-medioides**	**Decreto 2106 -2019**	**Resumen**
1	MEDELLIN	Especial	
2	BELLO	Tipo 1	
3	ENVIGADO	Tipo 1	
4	GIRARDOTA	Tipo 3	
5	GUARNE	Tipo 3	
6	ITAGUI	Tipo 1	
7	LA ESTRELLA	Tipo 2	
8	RIONEGRO	Tipo 1	
9	SABANETA	Tipo 1	
10	SONSON	Tipo 5	
11	BARRANQUILLA	Especial	
12	GALAPA	Tipo 5	
13	MALAMBO	Tipo 3	
14	SOLEDAD	Tipo 2	
15	BOGOTA, D.C.	Especial	
16	CARTAGENA DE INDIAS	Especial	
17	TUNJA	Tipo 1	
18	NOBSA	Tipo 5	
19	SOGAMOSO	Tipo 2	
20	MANIZALES	Tipo 1	
21	POPAYAN	Tipo 2	
22	CALOTO	Tipo 6	
23	VALLEDUPAR	Tipo 1	
24	MONTERIA	Tipo 1	
25	CAJICA	Tipo 2	
26	CHIA	Tipo 1	
27	COTA	Tipo 2	
28	FACATATIVA	Tipo 2	
29	FUNZA	Tipo 1	
30	MADRID	Tipo 2	
31	YOPAL	Tipo 2	
32	MOSQUERA	Tipo 1	
33	SOACHA	Tipo 1	
34	SOPO	Tipo 3	
35	TENJO	Tipo 3	
36	TOCANCIPA	Tipo 2	
37	ZIPAQUIRA	Tipo 2	
38	NEIVA	Tipo 1	
39	SANTA MARTA	Tipo 1	

Tipo de municipio	Cantidad	%
Especial	6	9.84%
Tipo 1	24	39.34%
Tipo 2	17	27.87%
Tipo 3	7	11.48%
Tipo 4	2	3.28%
Tipo 5	3	4.92%
Tipo 6	2	3.28%
Total general	61	100.00%

40	VILLAVICENCIO	Tipo 1	
41	PUERTO GAITAN	Tipo 3	
42	PASTO	Tipo 1	
43	IPIALES	Tipo 3	
44	CUCUTA	Tipo 1	
45	ARMENIA	Tipo 1	
46	PEREIRA	Tipo 1	
47	DOSQUEBRADAS	Tipo 2	
48	BUCARAMANGA	Especial	
49	BARRANCABERMEJA	Tipo 1	
50	GIRON	Tipo 2	
51	SINCELEJO	Tipo 2	
52	IBAGUE	Tipo 1	
53	ESPINAL	Tipo 4	
54	CALI	Especial	
55	BUENAVENTURA	Tipo 2	
56	GUADALAJARA DE BUGA	Tipo 2	
57	CARTAGO	Tipo 4	
58	PALMIRA	Tipo 1	
59	TULUA	Tipo 2	
60	YOTOCO	Tipo 6	
61	YUMBO	Tipo 1	

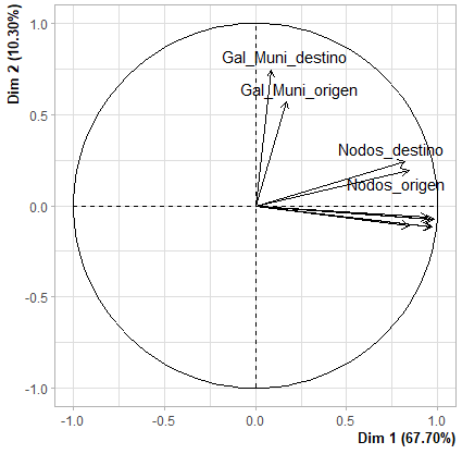
En la tabla anterior, se encuentran los municipios de acuerdo con el Decreto 2106 de 2019, de los municipios del tipo especial aparecen todos los seis municipios, de la lista de municipios de categoría 1 únicamente no aparece en la lista el municipio de Floridablanca en Santander. Buenaventura y Tocancipá los cuales son municipios de importancia estratégica para el movimiento de carga, en Buenaventura se encuentra el puerto más importante del país en términos de toneladas movilizadas y la ciudad de Tocancipá donde se encuentran importantes industrias de nivel nacional no se encuentran en las dos primeras categorías del decreto.

Se confirma la hipótesis planteada al inicio del documento, donde el Decreto 2106 de 2019 no cumplía con los criterios suficientes para ser utilizada en la clasificación de los municipios de acuerdo con el movimiento de carga.

Para la comparación de los valores medios de las variables se realiza un análisis de diferencia de medias entre los grupos de cada modelo, de esta forma, se busca comprobar la efectiva división de los grupos dentro de cada modelo.

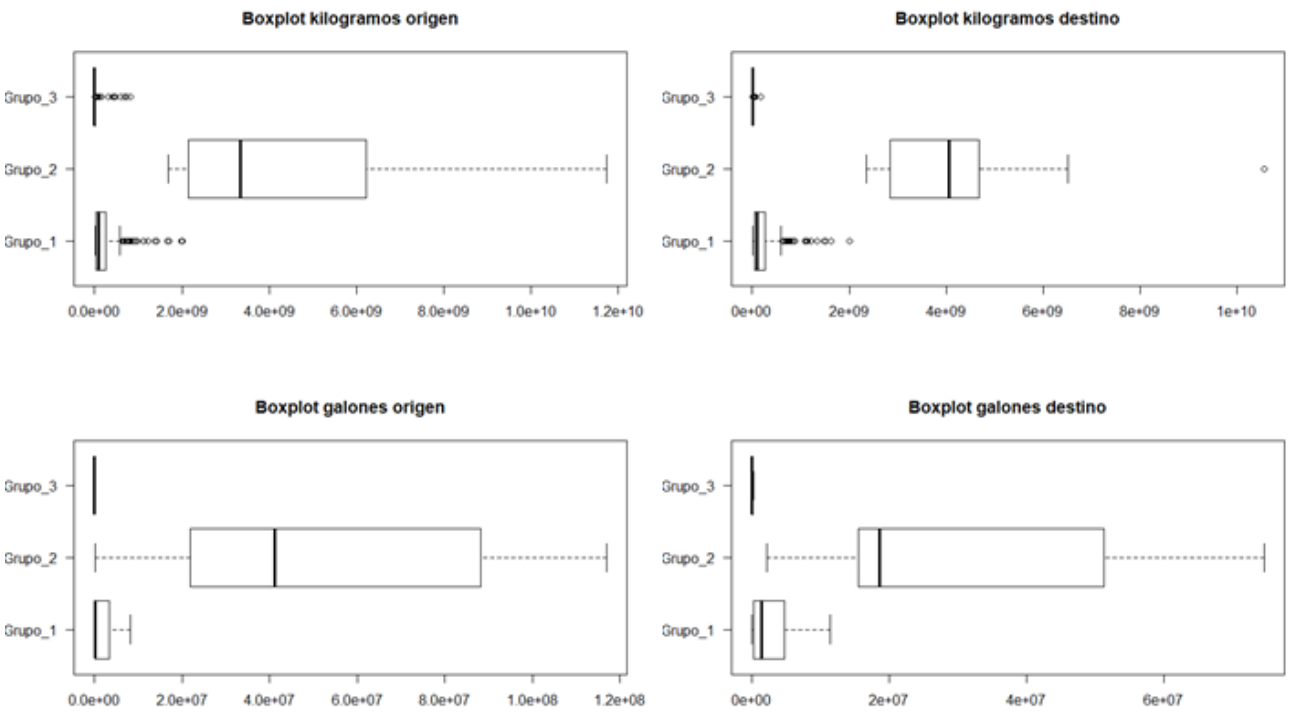
Se realizó un análisis de componentes principales y se identifican las variables que tienen mayor linealidad entre las variables y las que no.

Figura 18. Diagrama de radar de correlación de las variables.



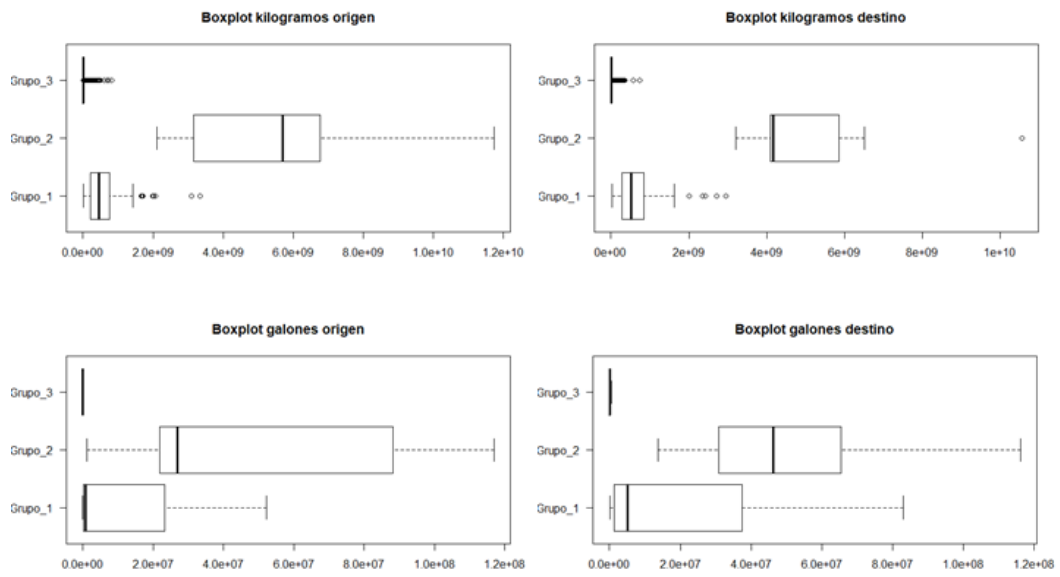
Se identifican las variables relacionadas con Galones y Kilogramos con las cuales se realizará la prueba para la igualdad de medias entre los grupos definidos.

Figura 19. Diagrama de caja y bigotes de modelo árbol de decisión Ward.D2



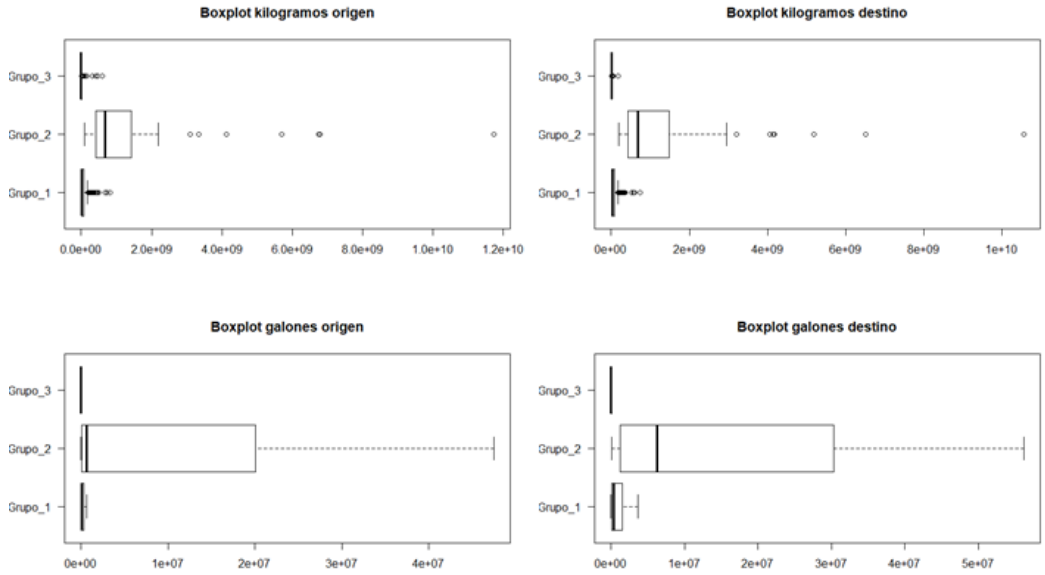
Se identifica en la figura que existen diferencias entre las medias de las variables analizadas, de modo tal que se recurre a la prueba de Wilcox para determinar la respectiva diferencia entre las variables analizadas y arrojan valores de prueba <0.05, por lo tanto, se rechaza la Ho, y se puede inferir que las variables no tienen una diferencia de medias igual a 0.

Figura 20. Diagrama de caja y bigotes de modelo k-medias



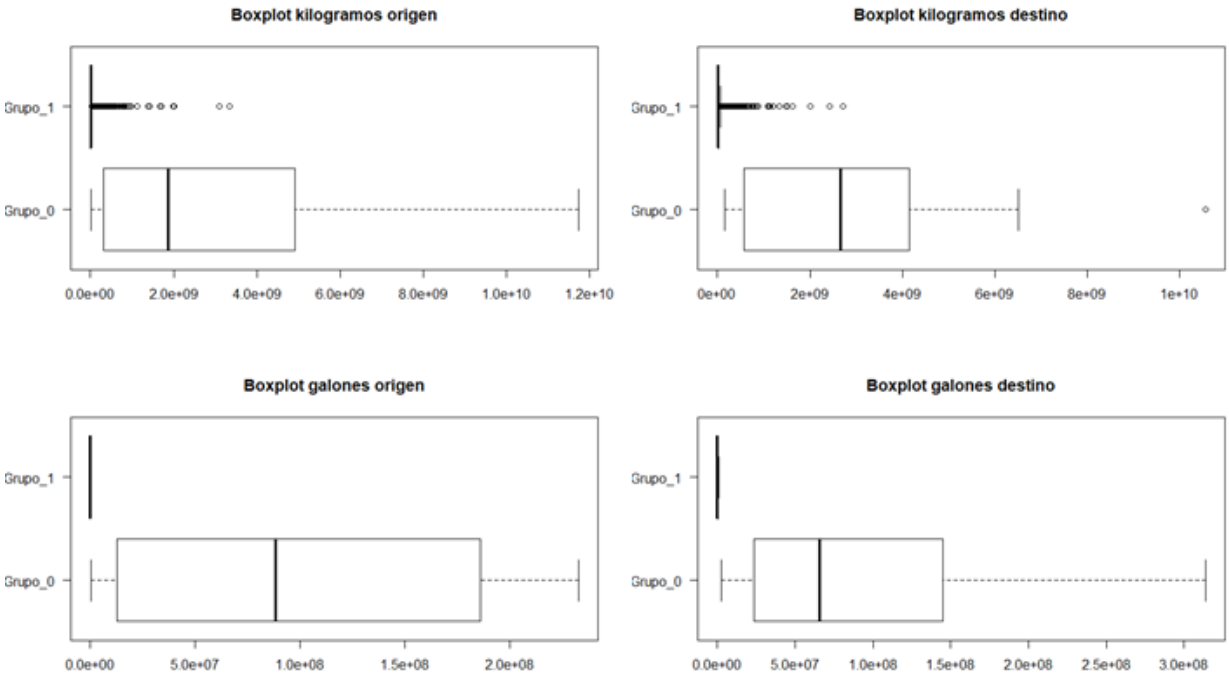
Se presenta en la figura, la posible existencia de diferencia entre las medias de las variables analizadas, lo cual se prueba a través del test de Wilcoxon, donde la hipótesis nula se refiere a que la mediana de los grupos es igual, versus, alguna de las medianas de los grupos es diferente. Con un 95% de confianza se concluye que las medianas son diferentes, confirmando el sentido de la existencia de las tres clasificaciones.

Figura 21. Diagrama de caja y bigotes de modelo k-medioides



En esta figura se observa el comportamiento relativamente similar entre los grupos 1 y 3 de las variables Galones Origen y Galones Destino, por lo que realizó el test Wilcoxon y con valores de prueba <0.05, se rechaza la hipotesis nula y se infiere que la diferencia entre las medias es diferente de 0.

Figura 22. Diagrama caja y bigotes de modelo DBSCAN



En esta figura, existe una diferencia marcada entre las variables analizadas, sin embargo, como se mencionó anteriormente, la utilidad del agrupamiento por encima de los 1000 nodos no es relevante. Se realiza el test de Wilcoxon para confirmar la diferencia y, por lo tanto, con un valor de prueba <0.05 se rechaza la hipótesis nula y se infiere que la diferencia entre las medias es diferente de 0.

De acuerdo con la evaluación de los modelos y el análisis de contexto, se busca tener homogeneidad en los miembros de los grupos, y heterogeneidad entre los grupos, esto se ha comprobado a través de las respectivas pruebas realizadas y finalmente con relación al contexto, el modelo que mejor apoya esta posición es el algoritmo de k-

medioides, teniendo una mejor distribución de municipios con vocación industrial y terminales portuarias.

Tabla 11. Criterios de evaluación de los modelos desarrollados.

Criterio	Árbol de clasificación	K-medias	K-Medioides	DBSCAN
Homogenidad entre individuos	218,9164	262,9018	157,6741	230,5654
Heterogenidad entre grupos	Cumple	Cumple	Cumple	Cumple
Distribución de individuos en grupos	7	8	6	9
Coeficiente variación	80,204	71,520	81,529	51,157

Se realiza una evaluación de los modelos teniendo en cuenta los criterios de homogeneidad entre individuos, heterogenidad entre grupos y distribución de los individuos en grupos, la homogeneidad entre individuos se determina sumando la desviación estándar de los tres grupos, la heterogeneidad se obtiene definiendo si las variables cumplen con la prueba de diferencia de medias, la distribución de individuos se evalúa con un ranking de clasificación de distribución entre los grupos generados y se determina el coeficiente de variación para medir el grado de dispersión, de esta forma el algoritmo k-medioides es elegido para definir el modelo.

4. DISCUSIÓN

El propósito de esta investigación es clasificar a los municipios del país, de acuerdo con sus características en el flujo de carga del servicio público de transporte, para soportar decisiones de política pública sobre el flujo de la carga en estos municipios.

Así las cosas, y partiendo de la caracterización del movimiento de carga representado en 10 variables para los 1100 municipios de Colombia para el año 2020, es posible la aplicación de técnicas de segmentación estadística para dar respuesta a la clasificación

buscada. Entre las múltiples técnicas de segmentación estadística están los algoritmos supervisados y no supervisados, dentro de los métodos seleccionados para obtener la clasificación fueron; k-medias, k-medioides, arboles de clasificación, modelo de densidad. En los modelos de k-medias, k-medioides y árboles de clasificación es necesario definir la cantidad de grupos a ser formados. El criterio de agrupación de k-medias es la mínima varianza entre los grupos formados, el criterio de k-medioides es la mediana de la distancia de los grupos respecto a un centroide y el árbol de clasificación es el mínimo de la distancia entre una función objetivo, en este caso el error de la suma de los cuadrados o varianza. de acuerdo con la utilización de estadísticos propios de cada algoritmo. De acuerdo con lo anterior, se podrían obtener resultados similares siempre que no existan datos atípicos, sin embargo, el conjunto de datos analizados la presencia de datos extremos que hacen parte de las mediciones naturales del fenómeno y por lo tanto no es adecuado excluirlos del análisis como sucede en algunas investigaciones, lo que implica que el método de k-medias y árboles de clasificación presentan un menor rendimiento respecto al método de k-medioides, toda vez que este ultimo se aplican criterios robustos como lo es la mediana.

Otra técnica utilizada para el objetivo planteado, fue el modelo de densidad discreto (DBSCAN), que a diferencia de los anteriores, no requiere el número de grupos a ser construidos como parámetro inicial, sino, la definición de un radio ε, que define la distancia máxima entre miembros de un grupo y un *minPts* que define la cantidad mínima de individuos para formar un grupo, lo que implica que el número óptimo de grupos a ser formados dependerá de la distribución espacial exhibida por el conjunto de datos, en la medida en que existan observaciones atípicas el radio definido no capturará la naturaleza

del fenómeno a clasificar, indicando que el número de grupos optimo es demasiado pequeño, como sucedió con el conjunto de datos analizados donde la sugerencia era armar un único grupo, donde los atípicos se consideraban ruido, resultados que desnaturaliza el objetivo de una clasificación en razón a que se busca identificar un grupo de individuos homogéneos al interior y heterogéneos al exterior, y no seguir trabajando con el único grupo inicial.

Del análisis anterior, y, teniendo en cuenta que en el conjunto de datos analizado existen comportamientos atípicos, se esperaba que el resultado del método de k-medioides fuera mejor que los otros tres métodos. Sin embargo, y debido a la naturaleza de estos métodos, no existen técnicas de prueba estadística que definan cual es la mejor clasificación, por lo cual se acudió a la aplicación de pruebas no paramétricas de comparación de medianas para ratificar la diferencia de los grupos construidos, así como la comparación de medidas de dispersión de estos grupos para aproximarse a una bondad de ajuste a la construcción de los grupos.

De acuerdo con lo expuesto, se propone que para la selección de modelos de clasificación se tengan presentes los criterios de homogeneidad entre los miembros de los grupos, la heterogeneidad intra grupal y la proporcionalidad de distribución de los miembros.

Se tiene que es posible realizar este tipo de clasificaciones utilizando herramientas de aprendizaje automático para diversos sectores de la economía de los mismos municipios. La aplicación de esta metodología permite ser ampliada para el movimiento de carga por tipos de productos, así como ser aplicada a otros modos de transporte. Finalmente, facilita la comprensión de la distribución geográfica de los sectores de mayor importancia en regiones específicas.

CONCLUSIONES

1. El algoritmo de k-medioides es el algoritmo seleccionado para la definición del modelo de clasificación y establecimiento de reglas de clasificación y vinculación de los municipios del país de acuerdo con el promedio de flujo de carga presentado durante un periodo de tiempo anual. Los valores de clasificación de municipios se presentan en sentido de llegadas y salidas, las variables definidas son viajes, kilogramos, galones, remesas y cantidad de orígenes y destinos.

Tabla 12. Valores promedio anuales de un municipio para asignación de grupo

	K-medioides			
Tipo	**Variable**	**Grupo 3**	**Grupo 1**	**Grupo 2**
	Viajes	221	4.178	108.613
	Kilogramos	4.515.868	67.668.739	1.362.297.595
	Galones	75.354	4.948.310	24.027.466
	Remesas	223	4.367	141.645
Llegada	Orígenes	10	82	422
	Viajes	403	5.796	97.722
	Kilogramos	3.740.153	68.272.752	1.368.048.894
	Galones	112.201	5.582.209	20.158.919
	Remesas	516	7.126	123.273
Salida	Destinos	21	89	260

2. Se confirma la hipótesis planteada al inicio del documento, donde el Decreto 2106 de 2019 no se adapta a los criterios para ser utilizada en la clasificación de los municipios de acuerdo con el movimiento de carga con la identificación de los tres grupos de municipios donde se evidencia la disimilitud de los movimientos de carga conforme a lo establecido en la tabla 12 donde se evidencia que todos los

municipios no son susceptibles de recibir las políticas aplicables respecto al servicio público de transporte de carga por carretera.

3. La importancia de analizar y de utilizar algoritmos de aprendizaje supervisado dentro del sector público para establecer reglas basadas en información soporta la toma de decisiones políticas que le den tratamiento a cada municipio conforme a sus características inherentes y no exista extralimitación u omisión en la exigencia de cumplimientos normativos emitidos por el Ministerio de Transporte.

REFERENCIAS BIBLIOGRÁFICAS

Aggarwal, C. (2015). *Data mining - The textbook.* New York: Springer Cham.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists.* Sebastopol: O'Reilly Media Inc.

Contaduría General de la Nación. (2021, 11 25). *Contaduria General de la Nación*. Retrieved from Categorización de Departamentos, distritos y municipios: https://www.contaduria.gov.co/categorizacion-de-departamentos-distritos-y-municipios

Departamento administrativo de la función pública. (2019, 11 22). *Decreto 2106 de 2019.* Retrieved from https://dapre.presidencia.gov.co/normativa/normativa/DECRETO%202106%20DEL%2022%20DE%20NOVIEMBRE%20DE%202019.pdf

Gonzalez-Marcos, A., & Alba-Elias, F. (2017). Machine learning en la industria: el caso de la siderurgia. *Economía industrial*, 55.63.

Guio Español, A. (2020). *Task force para el desarrollo e implementación de la inteligencia artificial en Colombia.* Bogotá: Consejería presidencial para asuntos económicos y transformación digital.

Igual, L., & Seguí, S. (2017). *Introduction to data science - A python approach to concepts, techniques and aplications.* New York: Springer Charm.

Ministerio de Transporte. (1996, 12 20). *Estatuto nacional de transporte.* Retrieved from Ley 336 de 1996: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=346

Ministerio de Transporte. (2015, 05 26). *Decreto único reglamentario del sector transporte.* Retrieved from Decreto 1079 de 2015: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=62514

Ministerio de Transporte. (2019, 03 14). Agenda logística. Bogotá, Colombia.

Ministerio de Transporte. (2021, 07 29). *Ministerio de Transporte*. Retrieved from ¿Quiénes somos?: https://www.mintransporte.gov.co/publicaciones/33/quienes_somos/

Ortiz Uribe, F. G., & Garcia Nieto, M. d. (2000). *Metodología de la investigación: El proceso y sus técnicas.* México: Limusa.

Pérez Borrero, I., & Gegúndez Arias, M. E. (2021). *Deep Learning. Fundamentos, teoría y aplicación.* Huelva: Universidad de Huelva.

R Core Team. (2021, 11 01). R: A language and environment for statistical computing. Viena, Austria: R Foundation for Statistical Computing.

República de Colombia. (1990, 07 20). *Secretaría del Senado*. Retrieved from Constitución Política de la República de Colombia: http://www.secretariasenado.gov.co/index.php/constitucion-politica

República de Colombia. (2019, 11 22). *Secretaria del senado*. Retrieved from Por el cual se dictan normas para simplificar, suprimir y reformar trámites, procesos y procedimientos innecesarios existentes en la administración pública.: http://www.secretariasenado.gov.co/senado/basedoc/decreto_2106_2019.html#TÍTULO%20I