MEJORA DE LA CAPACIDAD DEL PROCESO OPERATIVO DE UNA EMPRESA DE RECOLECCIÓN DE RESIDUOS SÓLIDOS URBANOS, INTEGRANDO *LEAN SIX SIGMA* Y CIENCIA DE DATOS

Mario Andrés Valencia Díaz

UNIVERSIDAD AUTÓNOMA DE MANIZALES

FACULTAD DE INGENIERÍA

MAESTRÍA EN INGENIERÍA

MANIZALES, COLOMBIA

2022

MEJORA DE LA CAPACIDAD DEL PROCESO OPERATIVO DE UNA EMPRESA DE RECOLECCIÓN DE RESIDUOS SÓLIDOS URBANOS, INTEGRANDO *LEAN SIX SIGMA* Y CIENCIA DE DATOS

AUTOR

MARIO ANDRÉS VALENCIA DÍAZ

DIRECTOR (A):

PH.D Reinel Tabares Soto

CODIRECTOR:

M.SC Cristian Felipe Jiménez Varón

PH.D Simón Orozco Arias

UNIVERSIDAD AUTÓNOMA DE MANIZALES

FACULTAD DE INGENIERÍA

MAESTRÍA EN INGENIERÍA

MANIZALES, COLOMBIA

2022

DEDICATORIA

Después de muchos años de estar fuera de las aulas, regresar a ellas e ingresar a una maestríaen investigación, implica un reto desde lo personal, familiar y laboral. Pero más allá de las implicaciones personales y laborales, esta nueva estrategia organizacional que hoy se evidencia en una tesis de maestría, requiere un apoyo y una credibilidad de parte del equipo de socios de SIGMA Ingeniería, para quienes está dedicado este primer trabajo en investigación.

Ellos… (Claudia, Caliche, Natalia, Juanis) se atrevieron a permitirnos explorar el camino derelacionamiento con la universidad en medio de la incertidumbre sobre el destino de este proceso, aferrados sólo a la credibilidad y la confianza que siempre me han brindado honestamente

AGRADECIMIENTOS

“En el proceso de materializar este trabajo de maestría, fue crucial encontrarnos en el camino con Alejandra María Restrepo Franco, (hoy una de nuestras socias en GeoStrategy) a ella todo el agradecimiento por su facilidad de ordenar el camino de una investigación y servir de guía cuando la cotidianidad de dirigir una compañía hacía perder el camino de la investigación. Al equipo de trabajo de GeoStrategy porque sus aportes dentro de la investigación nos permitieron avanzar en el conocimiento y lograr los objetivos planteados.

Por último, agradecer a la Universidad Autónoma de Manizales y a su rector por facilitar sus sagrados espacios de conocimiento, para experimentar una nueva forma de relacionamiento entre la universidad y la empresa, que pretende unificar el lenguaje y valorar todos los esfuerzos que se realizan para hacer ciencia aplicada a la región.

RESUMEN

Éste trabajo de investigación determina las acciones de mejora en la capacidad del proceso derecolección de residuos sólidos urbanos integrando *Lean Six Sigma* y Ciencia de Datos, para un caso de estudio en la ciudad de Manizales Caldas; lo anterior, con el fin de optimizar procesos logísticos en la recolección de residuos sólidos urbanos. Este proyecto partió de un enfoque cuantitativo no experimental, iniciando con la extracción de los datos del sistema de información GEOASEO, identificando y caracterizando las variables del modelo de datos hastael entrenamiento de modelos que permiten la predicción del tiempo y las toneladas recogidas del proceso, como variables de calidad requeridas para este modelo de negocio. Esta investigación es aplicada a la industria de recolección de las ciudades, utilizando herramientas de*Lean Six Sigma* innovando su análisis al incluir la ciencia de datos como elemento estratégicoen la selección de la relación entre las variables y el entrenamiento de los modelos que permitenpredecir el comportamiento del sistema en su estado actual; además integró el resultado de la capacidad del proceso, los sistemas de gestión actuales en el caso de estudio. En este sentido, esta investigación se permite argumentar la necesidad de ajustar las planeaciones estratégicas por mediodel análisis histórico de los datos y de máquinas de aprendizaje que retroalimentan lascondiciones cambiantes del modelo. La solución presentada permite dar respuesta a la necesidad específica de optimizar el proceso de recolección de residuos sólidos urbanos, para este caso, en la cuidad de Manizales.

Palabras Clave: Ciudades inteligentes, Lean Six Sigma, Ciencia de Datos, Residuos Sólidos,GEOASEO by Sigma Ingeniería S. A

ABSTRACT

This research work determines the actions to improve the capacity of the urban solid waste collection process by integrating Lean Six Sigma and Data Science, for a case study in the city of Manizales Caldas; the above, in order to optimize logistical processes in the collection of urban solid waste. This project started from a non-experimental quantitative approach, starting with the extraction of data from the GEOASEO information system, identifying and characterizing the variables of the data model until the training of models that allow the prediction of time and tons collected from the process, as quality variables required for this business model. This research is applied to the city collection industry, using Lean Six Sigma tools, innovating its analysis by including data science as a strategic element in the selection of the relationship between the variables and the training of the models that allow predicting the behavior of the system in its current state; It also integrated the result of the process capacity, the current management systems in the case study. In this sense, this research allows us to argue the need to adjust strategic planning through historical data analysis and learning machines that feed back the changing conditions of the model. The solution presented allows us to respond to the specific need to optimize the urban solid waste collection process, in this case, in the city of Manizales.

Keywords: Smart Cities, Lean Six Sigma, Data Science, Solid Waste, GEOASEO by Sigma Ingeniería S. A

CONTENIDO

LISTA DE SIMBOLOS Y ABREVIATURAS

ABREVIATURA	TÉRMINO
LSS	*Lean Six Sigma*
DMAIC	Definir, Medir, Analizar, Mejorar, Controlar
SIG	Sistema de Información Geográfico
RRS	Recolección de residuos sólidos
RRSU	Recolección de residuos sólidos urbanos
GPS	Sistemas de Posicionamiento Global
IA	Inteligencia Artificial
TIC	Tecnologías de Información y Comunicación
IOT	Internet de las cosas
CRA	Comisión de Regulación de Agua Potable y Saneamiento
RFID	Identificación por Radio Frecuencia
ODS	Objetivos de Desarrollo Sostenible
SEDESOL	Secretaría de Desarrollo Social de México
OCDE	Organización para la Cooperación y el Desarrollo Económico
DNP	Departamento Nacional de Planeación
CEPAL	Comisión Económica para América Latina y el Caribe
RNA	Redes Neuronales Artificiales
MSE	Error Cuadrático Medio
RMSE	Raíz Cuadrada del Error Cuadrático Medio
MAE	Error Medio Absoluto
LSTM	Red neuronal de memoria a corto y largo plazo
BLSTM	Redes LSTM bidireccionales
R^2	Coeficiente de Determinación
VRP	Problemas de Enrutamiento de Vehículos
CP	Índice de capacidad del proceso
CPK	Indicador de capacidad real
LIE	Límite inferior de calidad del proceso

LSE	Límite superior de la calidad del proceso
PCA	Análisis de Componentes Principales
FNN	Fully Connected Neural Network
TRL	Nivel de Madurez Tecnológica
ASI	Análisis del sistema
DSI	Diseño del sistema
IAS	Implantación
RELU	Función de activación unidad lineal rectificada
N	Número de Neuronas

LISTA DE TABLAS

LISTA DE FIGURAS

1 PRESENTACIÓN

Cada vez que se inicia el camino de la mejora continua, se evidencia la necesidad de definir el conjunto de variables que describen el problema para caracterizarlas, relacionarlas y poder explicar el comportamiento de los sucesos que se desencadenan al interactuar de una forma específica.

En el caso de las operaciones de recolección de residuos y aseo, las diferentes variables generan grandes volúmenes de datos cuantitativos que describen, bajo una visión retrospectiva, como fue el comportamiento de dicha operación en ese instante de tiempo. En este sentido, poder determinar los patrones de comportamiento entre las variables (dependientes e independientes) que explican el desempeño de la operación, pueden dar paso a análisis predictivos y prescriptivos como antesala a los procesos de mejora continua, así entonces la ciencia de datos se convierte en uno de elementos que estimulan el desarrollo de esta investigación.

La mejora continua, apunta al incremento de la productividad, disminuyendo los desperdicios que no arrojan valor para el cliente e incrementando la capacidad del proceso, en consecuencia, la mejora continua, permite establecer estrategias sistemáticas de mejoramiento en busca de la excelencia operacional (Felizzola Jiménez & Luna Amaya, 2014) . No basta con definir de forma clara el problema, identificar los impactos, medir correctamente las variables, relacionarlas y determinar el grado de dependencia entre las mismas, es necesario además buscar y evidenciar las relaciones que existen entre ellas y que sirve como insumo para el diagnóstico de la operación.

Por tanto, en la presente investigación se estudió el desarrollo de un proceso de mejora continua bajo el modelo de *Lean Six Sigm*a (LSS), integrando la metodología de desarrollo DMAIC de mejora de proceso (Definir, Medir, Analizar, Mejorar, Controlar) y las técnicas de análisis de datos (Garza Ríos et al., 2016), lo cual permite articular lo mejor de estos dos mundos a los procesos de mejora continua a través del desarrollo de modelos predictivos capaces de predecir el comportamiento del sistema a partir de las variables sugeridas por el modelo, esto con el fin de optimizar el proceso de recolección de residuos sólidos urbanos en este caso, para la ciudad de Manizales.

Cabe destacar, que este proyecto se ha articulado con la empresa SIGMA INGENIERÍA, la cual cuenta con una herramienta tecnológica Geoaseo que logra captar información de las empresas de aseo en operaciones como gestión de rutas, recolección, barrido, atención de aforos y servicios especiales, seguimiento vehicular entre otras variables que se presentan en el proceso de aseo de la ciudad (SIGMA INGENIERIA S.A, n.d.).

Geoaseo es un sistema de información basado en SIG (Sistema de Información Geográfico), con entornos web y móvil, que posibilita la gestión de los procesos operativos, capturando y procesando datos geográficos y alfanuméricos generando información, alineados a los indicadores de gestión de cada empresa. La plataforma concentra sus soluciones en mejorar la calidad, eficacia, productividad y rentabilidad de la organización. En este sentido, esta propuesta pretende gestionar eficientemente las diferentes operaciones y generar conocimiento que apunte a los objetivos estratégicos de la organización a partir de la ciencia de datos.

Los resultados principales de la investigación se basaron inicialmente en la identificación de las especificaciones de calidad de la operación de RRS las cuales dieron línea a poder identificar las variables internas del modelo de recolección de residuos y medir mediante la capacidad del proceso el impacto sobre las especificaciones de calidad del modelo de negocio. Finalmente, se analizó, diseñó e implementó un prototipo en TRL6 donde se encuentran integrados 2 modelos basados en técnicas de ciencia de datos específicamente en Machine Learning para la predicción de las variables estratégicas que definen la capacidad de la operación de RRS, los modelos implantados permiten predecir las nuevas planeaciones estratégicas y mejorar la capacidad del proceso , además podrían ser utilizadas en línea y poder determinar en un momento específico de la operación cuál va a ser el comportamiento de la ruta a partir de las condiciones como el día, kilómetros, tiempo y número de viajes. La solución presentada permite dar respuesta a la necesidad específica de procesar información a través de los módulos que actualmente se tienen configurados de recolección de residuos sólidos urbanos e identificar las acciones de mejora que podría tener la operación.

2 ANTECEDENTES

Para dar respuesta a las necesidades de mejora hace falta iniciar con la identificación y caracterización de las variables involucradas en el proceso, además de tener conocimiento en torno a la industria de aseo y recolección de residuos sólidos, ya que esta es una operación que cuenta con muchas dimensiones las cuales han sido objeto de estudio en los últimos años: la logística de recolección apoyada en SIG (Ogryzek & Wolny-Kucí, 2021) en sus diferentes modelos de negocios (residuos ordinarios, peligrosos, industriales, reciclaje).

En particular, son varios los enfoques en la búsqueda de optimizar los procesos logísticos en la recolección de residuos sólidos (RRS). Ya que cuando hablamos de estas estrategias en los últimos años se han desarrollo entorno a darle una visión mucho más sustentable del proceso y es que no es posible desestimar el impacto que hoy en día tiene este aspecto, como es el caso de (Hannan et al., 2020), donde se expone una revisión integral de cómo el desafío de la gestión de los residuos es el aumento demográfico y urbanístico que se da de manera desproporcionada, lo cual conlleva a los procesos de RRS a dar soluciones desde la optimización, este artículo de revisión propone una revisión completa de todo lo que este proceso requiere y evidencia las diferentes investigaciones en torno a la misma como es el caso de los objetivos de optimización en RRS, ya que son varios los puntos de concentración en torno a esto, incluyendo la optimización de rutas, tiempo, localización/relocalización de los puntos de recogida, optimización de los contenedores, costos, estaciones de transferencia, recolección e impactos ambientales, así como también muestra las restricciones existentes en torno a capacidad, demanda, balance en masa, tiempo, tipo de desechos, medioambientales, regulatorios, políticos y sociales además se destacan los métodos de optimización disponible clasificándolos en diferentes categorías y tecnologías como es el caso de los enfoques deterministas y estocásticos en los que se destacan los métodos probabilísticos y lógico-difusos. Donde se plantea que todas las soluciones en un primer momento partieron de soluciones convencionales para luego integrar las perspectivas heurísticas y meta-heurísticas debido a la mayor complejidad del problema de recolección y el tiempo computacional.

A través del tiempo, se han venido también integrando nuevas tecnologías basadas en IoT, que han aportado al desarrollo tecnológico en los procesos de RRS, como punto en común se ha dado la integración de los SIG para los modelados y en la actualidad tecnologías más avanzadas tales como Identificación por Radio Frecuencia (RFID, por sus siglas en inglés), los Sistemas de Posicionamiento Global (GPS, por sus siglas en inglés), entre otros, junto a los contenedores inteligentes y diferentes sensores. Para entender más a fondo los impactos que tiene el uso de tecnologías avanzadas y los software SIG que se exponen en (Hannan et al., 2020), en la optimización del proceso, nos encontramos a (Betanzo-Quezada et al., 2016) en el marco de la gestión en logística y seguimiento de 71 rutas del proceso recolección de residuos sólidos urbanos, se expone un caso de estudio de la evaluación del sistema de recolección en el municipio de Querétaro en México en base a los datos obtenidos en el monitoreo de los vehículos con dispositivos GPS, dicha evaluación se elaboró determinando la variación de los recorridos planeados contra los obtenidos por medio del dispositivo, teniendo en cuenta los costos asociados y condiciones prevalecientes en dicha operación, este procedimiento solo se realizó para residuos sólidos urbanos, excluyendo los de tipo industrial y comercial, tal y como se plantea en el marco del trabajo ejecutado en esta investigación.

Asimismo, (Betanzo-Quezada et al., 2016) detalla que en el contexto mexicano se revela un incremento en la generación de basuras a 0.90 kg/hab/día en 2004 a una proyección estimada de 1.06 kg/hab/día al año 2020, datos de la Secretaría de Desarrollo Social de México (SEDESOL), además que la generación de residuos sólidos urbanos per cápita es uno de los indicadores para el desarrollo sustentable tal y como se muestra en la Organización para la Cooperación y el Desarrollo Económico (OCDE). Para realizar el estudio comparativo se apoyó la investigación de la construcción de tableros de control, asimismo se realizó una discusión en contexto a los cambios realizados desde la empresa recolectora durante el curso de la investigación y se expone sobre la confiabilidad de los resultados obtenidos, los parámetros empleados en el diseño de las rutas y los aspectos normativos y de planeación, concluyendo que la inclusión de dichos dispositivos a la ejecución de dicha operación demuestran ser de gran utilidad, que además la combinación entre dichos dispositivo y una adecuada sistema de procesamiento de los datos, aportan una herramienta de bajo costo para el

monitoreo, se considera además evidente a partir de esta investigación la variación que existe entre las rutas planeadas y su ejecución real. Y que dicha información es oportuna y confiable para realizar los cambios necesarios de manera eficiente con respecto a las cambiantes necesidades de las ciudades.

Como se expuso anteriormente la importancia de poder obtener información de manera confiable a través de los SIG y de las tecnologías avanzadas como GPS, es tener a disposición gran cantidad de información que relate lo que ocurre en la operación mes tras mes, de la recolección de esta información en años los métodos matemático-estadísticos generan la posibilidad de realizar pronósticos, como es el caso de (Sodanil & Chatthong, 2014), en donde se muestran los modelos de pronóstico para residuos sólidos como método efectivo para la administración y planeación del proceso. Teniendo en cuenta que los principales objetivos del análisis por serie de tiempos son, por un lado, la identificación de la naturaleza de los fenómenos representados por la secuenciación de las observaciones, y, por otro lado, el pronóstico prediciendo futuros valores del análisis de las variables en series de tiempo. Los trabajos relacionados en torno al pronóstico por series de tiempo están relacionados a Suavizado Exponencial, Cajas-Jenkins, Redes Neuronales Artificiales, Método ARIMA y Redes Neuronales. En particular el uso de Redes Neuronales de Perceptrón Multicapa.

Asimismo, este artículo ubica como caso de estudio a la ciudad de Bangkok capital de Tailandia donde se recolectó la información entre Octubre del 2002 y Julio del 2013, para un total de 130 meses, esta información se recolectó de las variables involucradas en la recolección bajo las directrices del Departamento de Medio Ambiente perteneciente bajo la Administración Metropolitana de Bangkok, posterior a esto se realizó un preprocesamiento de la información a través de normalización tipo Z-Score para evaluar puntos atípicos y desviaciones de medias para conjuntos de 12 meses, posteriormente se realiza una rutina de 3 pasos para los correctos diagnósticos debido a que los datos no son estacionarios, finalmente se realiza en entrenamiento de Redes Neuronales Artificiales (RNA) con ayuda de una herramienta llamada RapidMiner (RapidMiner, n.d.), dividiendo el conjunto de datos en 2 subconjuntos de los cuales el 80% se utiliza para el entrenamiento y el 20% se utiliza para las pruebas y aplicando un perceptrón de 3 capas, 1 unidad de capa ocultas con 35 y 1 salida. Posteriormente se

realizó la evaluación del modelo a través de las métricas Error Cuadrático Medio (MSE, por sus siglas en inglés), precisión y el coeficiente de determinación R^2. Teniendo en cuenta que el modelo de RNA fue entrenado con un algoritmo de retropropagación. Los mejores resultados mostraron que una estructura de red neuronal de 3-35-1 presenta el mayor rendimiento con una precisión de predicción de 0,870 y un MSE de 0,2333.

En este mismo sentido de realizar un correcto pronóstico de la generación de los RSU, de manera más complejas se encuentra el trabajo realizado de (Jammeli et al., 2021), el cual integra modelos secuenciales de regresión/clasificación en red neuronal de memoria a corto y largo plazo (LSTM, por sus siglas en inglés) y las redes LSTM bidireccionales (BLSTM) para así modelar datos temporales de generación de residuos, recopilados durante un año en distintas zonas de la ciudad de Sousse, Túnez, en aras de predecir el número necesarios de depósitos (llamados también contenedores), para una correcta gestión de los esfuerzos realizados apoyados en el marco de desarrollo de las ciudades inteligentes. Este artículo además realiza una completa evaluación experimental y una comparación de los métodos secuenciales de regresión/clasificación en contraste a los métodos habituales de represión/clasificación no secuenciales relevantes. Como resultados experimentales a la utilización de estos métodos secuenciales, se demuestra la superioridad de la regresión/clasificación LSTM y BLSTM en términos del desempeño del pronóstico realizado en relación con el número de contenedores, asimismo detalla las fortalezas y debilidades de ambos métodos, evidencia la importancia de la consideración de los datos temporales para realizar pronósticos de la generación de los residuos sólidos.

En el marco de referencia de la formulación del proyecto, se considera relevante además de los 4 artículos presentados, los artículos (Zhang et al., 2021) y (Akbarpour et al., 2021), en donde (Zhang et al., 2021) donde se detalla de manera específica una revisión de lo que abarca los Problemas de Enrutamiento de Vehículos (VRP, por su siglas en inglés), que clasifica de acuerdo a sus características y aplicaciones prácticas de estos tipo de problemas, también muestra los modelos y algoritmos implementados en sus soluciones. Por su parte, (Akbarpour et al., 2021) propone un innovador sistema de gestión de residuos en el marco de las ciudades inteligente, por medio de optimización estocástica utilizando el problema de enrutamiento de vehículos, integrando la

minimización del costo total de transporte y la maximización de los ingresos por elementos reciclados, teniendo en cuenta los esfuerzos dados en torno a políticas medioambientales más sustentables.

Por otro lado, es indispensable tener en cuenta el panorama de los estudios realizados para diferentes zonas, ya que es un proceso que como se evidencia, requiere del conocimiento de los datos en territorio para que sea eficiente y óptimo además de articularse con la mejora continua, tema central del proyecto de investigación. En la Tabla 1, se evidencia la consolidación de algunos antecedentes de los procesos aplicados a territorio y el impacto que dichos antecedentes tienen sobre esta investigación.

Tabla 1. Antecedentes Asociados A La Industria

No.	Antecedente	Impacto
1	Disposición Final de Residuos Sólidos Informe Nacional– 2019 (Superservicios, 2019)	El informe de gestión de la Superintendencia de servicios públicos domiciliarios del 2019 da una visión general de cómo se articula la industria de servicios públicos en Colombia y establece hechos y datos sobre el impacto que puede llegar a tener estos proyectos como investigación aplicada, articulando a los indicadores estratégicos del sector.
2	Localización del punto óptimo de partida en el problema de ruteo vehicular con capacidad restringida (CVRP) (Soto Mejía et al., 2019)	Esta investigación realizada en Dosquebradas, Risaralda (Colombia), hace uso del algoritmo genéticos modificado Chu Beasley, para disminuir los tiempos de ruta, aprovechando la clusterización antes del ruteo de los vehículos, es muy importante para esta investigación porque parte de las optimizaciones del proceso pueden darse en los diseños de las rutas.
3	Brecha del servicio de limpieza pública en la ciudad de Tingo María, Perú (Montalvo, 2018)	Este estudio realizado en el 2018, en la ciudad de Tingo María, (Perú), permite tener una óptica de las condiciones que los usuarios de un servicio público de aseo pueden tener como percepción de calidad de un servicio. Al hacer uso de LSS, la metodología DMIAC, en su primera etapa de definición, procura establecer cuáles son las variables que el usuario final determina como calidad para lo cual este estudio puede ser un referente importante al analizar variables como: “unidades móviles del sistema de recolección de residuos sólidos, tipo de infraestructura de manejo y tratamiento de residuos sólidos, nivel de cumplimiento de responsabilidades adquiridas por el área de limpieza pública, nivel de interés en solucionar el problema del usuario, la ejecución del servicio según responsabilidad adquirida, el énfasis en registrar los errores, la exactitud en la comunicación de la realización de los servicios, disposición a ayudar de los trabajadores, disposición de los trabajadores a responder preguntas, manejo de residuos sólidos” entre otros.
4	El paradigma de las Smart Cities en el marco de la gobernanza urbana (Tarín, 2018)	Este estudio del 2018 enfoca desde una metodología descriptiva deductiva, el papel de la gobernanza dentro del concepto de las ciudades inteligentes, y es importante para este estudio, porque nos establece los elementos sobre los cuales las smart cities, están alterando la forma de relacionarse con los servicios públicos, retos como “futuros de la participación, la ética pública, la sostenibilidad, la eficacia y eficiencia, la transparencia y la rendición de cuentas”, deben ser tenidos en cuenta al momento de establecer los elementos de la calidad del servicio en la empresas de servicios públicos de aseo, actuales y futuras.

5	Gestión de residuos sólidos domiciliarios en la ciudad de Villavicencio. Una mirada desde los grupos de interés: Empresa, estado y comunidad (Trujillo González et al., 2017)	El estudio de la Universidad de Caldas del 2017, sobre las empresas de residuos sólidos domiciliarios en la ciudad de Villavicencio, presenta una mirada desde los grupos de interés que nos permite tener una perspectiva más nacional de los que esperan los grupos de interés de este tipo de servicio y así por sustentar los criterios de calidad del servicio. Es una de las primeras etapas de la metodología y la actualidad del estudio nos brinda una visión desde los tres grupos de interés: Empresa, Estado y Comunidad.
6	Metaheurística para la solución del "Problema de diseño de la red de transporte" multiobjetivo con demanda multiperiodo (Garzón et al., 2017)	A pesar de que, el presente estudio no parece tener una conexión directa con nuestra industria de aseo, el problema de diseñar redes de tránsito con múltiples objetivos puede convertirse rápidamente en un elemento de optimización o de análisis del comportamiento de nuestro sistema, y el tema abordado en esta investigación incluye búsqueda de vecindades cambiantes (VNC) en condiciones de mañana, tarde y noche, arrojando muy buenos resultados en las pruebas iniciales.
7	Estudio exploratorio en torno a las potencialidades de los recicladores de oficio para la construcción de nueva política pública con inclusión social en el sistema de aseo en Bogotá D. C. (Serrano, 2016)	Este estudio cuenta con la particularidad de analizar un sistema de recolección desde la perspectiva del reciclaje y sus impactos en los grupos sociales. Es importante no perder de vista este referente porque nos puede indicar condiciones propias del servicio desde un actor y un modelo de negocio distinto, pero que se articula a nuestro objeto de estudio. Además, permite contar con una nueva lectura desde una operación de aseo en un territorio como Bogotá, enriqueciendo la discusión y entregando nuevas dimensiones del modelo de negocio.
8	Impacto de los líderes en la productividad de las empresas de servicio de aseo en la ciudad de Barranquilla (Rivera Cerpa & Conrado Tobón, 2016)	Este estudio, con un corte más gerencial, centra sus discusiones alrededor de la productividad, pero además nos permite tener un panorama de la industria de aseo, desde otra región de Colombia (Barranquilla). Realiza una revisión de la literatura entre 1985 y 2015, e identifica el clima organizacional, la capacidad de liderazgo o el estilo de liderazgo, como detonante del cumplimiento de los objetivos y por consiguiente de la productividad de la industria.
9	Un enfoque híbrido de agrupamiento y optimización entera mixta para el problema de servicios de recolección selectiva de residuos sólidos domésticos (Patiño Chirva et al., 2016)	De nuevo tenemos análisis de las operaciones de aseo desde la perspectiva de la logística, específicamente desde la optimización de las rutas, en este caso la ciudad de estudio es Bogotá, y las consideraciones del modelo incluyen variables como capacidad, duración máxima de la jornada, condiciones normativas del modelo en Bogotá, y hace uso de un paquete comercial para resolver las rutas selectivas de forma óptima.
10	Mejora del servicio de recolección de residuos sólidos urbanos empleando herramientas SIG: un caso de estudio (Araiza Aguilar & José Zambrano, 2015)	El estudio se realiza en México, en el 2015, pero pone en manifiesto los impactos sobre el medio ambiente y la productividad, cuando se realizan procesos de recolección ineficientes, el estudio indica entre un 50% y 90% de incremento en los costos por una mala gestión del servicio de aseo en su etapa de recolección. Más allá de las connotaciones de la forma y el diseño de las operaciones en México, este estudio es especialmente importante porque nos entrega elementos de medición de la productividad.
11	Factores Clave en la Gestión de Tecnología de Información para Sistemas de Gobierno Inteligente (Góngora & Bernal, 2015).	Las empresas de servicios públicos de aseo están inmersas dentro de una arquitectura de ciudades inteligentes y están prestas a apropiar y suministrar información para la gobernanza de las ciudades inteligentes, por esa razón hemos referenciado este antecedente porque es necesario articular a los conceptos y las exigencias de las "Smart Cities", y este artículo en particular ofrece factores claves en la gestión de las TIC para los gobiernos inteligentes.
12	Factores críticos para la medición de la calidad del servicio del aseo urbano en el municipio Maracaibo (Sáez, 2011)	A pesar de la distancia en el tiempo (2011) del presente artículo, entrega una visión de un servicio público de recolección en un país como Venezuela y un municipio como Maracaibo. El interés sobre este antecedente es su metodología y los elementos claves que determina para la etapa de definir del DMAIC; los elementos claves de la calidad del servicio de aseo, pueden ser un insumo para la definición de la calidad del servicio esperado por los clientes de la industria en Colombia.

Elaboración Propia

Como se evidencia a pesar de contar con investigaciones suficientes alrededor de la industria, la mayoría de ellas se concentra en la optimización de las rutas, y la determinación de los elementos de calidad del servicio, es importante en este punto hacer claridad, que el presente estudio pretende analizar los datos de las operaciones y determinar sus puntos de mejora por medio de la metodología LSS y la ciencia de los datos. La articulación con las metodologías de optimización de procesos no ha sido altamente explorada para las industrias de servicios públicos de aseo. Lo cual genera la necesidad de establecer en qué industrias y en qué áreas del conocimiento, las metodologías de mejora continua como LSS han sido utilizadas, con el objetivo de usar dichos estudios como punto de partida para determinar el impacto que esta metodología tiene sobre las industrias, por lo tanto, en la Tabla 2 vamos a consolidar los antecedentes sobre la metodología LSS.

Tabla 2. Antecedentes Asociados A Lean Six Sigma

No.	**Antecedente**	**Sector**	**Impacto**
1	Propuesta para implementar *Lean Six Sigma* en el departamento de servicio al cliente en una empresa del sector retail (Albañil & Martínez, 2019)	Servicios de Retail	La variable resultante de la calidad de este tipo de industria termina siendo el tiempo de espera, partiendo de un 42% de incumplimiento se logran mejoras significativas al proceso. Para nuestro estudio la aplicación de la metodología hace de este estudio algo interesante en su análisis de los tiempos.
2	Ciclo DMAIC en Latinoamérica: Análisis de la relación con el Producto Interno Bruto aplicación (Rojas Salazar & Pérez Olguín, 2019)	Gobierno	Según las conclusiones del estudio, a mayor PIB, mayor número de aplicaciones de DMAIC hay en los países, lo que infiere un incremento en la productividad
3	The revolution Lean Six Sigma 4.0 (Arcidiacono & Pieroni, 2018)	Servicios de Salud	Cada vez más LSS toma relevancia en las industrias del servicio, en este caso el estudio enuncia las fortalezas que tienen los datos para la mejora de procesos y las técnicas que a partir de la ciencia de los datos podemos usar para facilitar y mejorar la aplicación de la metodología.

4	Sistemas De Producción Competitivos Mediante La Implementación De La Herramienta *Lean Manufacturing* (Vargas-Hernández et al., 2018)	Manufactura	Uno de los artículos más claros en términos de la relación LSS y la ciencia de los datos, imprescindible para entender cuándo y dónde hacer uso de los recursos de estas dos líneas de trabajo
5	Proyectos de desarrollo de proveedores que usan *Six Sigma*: un análisis de caso en Schneider Electric Colombia S.A (Muñoz, 2018)	Productos de Energía	Este trabajo se considera un proyecto de desarrollo de proveedores.
6	Mejora de Procesos ERP´s (Enterprise Resource Planning) con *Lean Six Sigma* (Alvarado Chávez, 2018)	Administrativos	Este proyecto impacta los flujos de datos de los productos entre Estados Unidos y México, logrando reducciones de 33% en menos de 3 meses y más de 40 mejoras significativas al proceso.
7	Enfoque seis sigma y proceso analítico jerárquico en empresa del sector lácteo (Herrera Vidal et al., 2017)	Sector de Lácteos	Su impacto se centra en aumentar la capacidad del proceso reduciendo los productos no conformes, por medio de six sigma en híbrido con análisis jerárquico, logrando acoplar las dos. metodologías para el logro de los objetivos
8	Modelo metodológico de implementación de *lean manufacturing* (Sarria Yépez et al., 2017)	Manufactura	Integra la metodología ICOM con lean manufacturing, con el objetivo de lograr una mejor apropiación y facilidad de implementación específicamente en la pequeña empresa
9	The AMSE *Lean Six Sigma* governance model (Arcidiacono et al., 2016)	Modelo de gobierno	Un trabajo muy interesante para garantizar la sostenibilidad de las implementaciones LSS, pero en este caso no será objeto de uso, porque el problema no incursiona en las etapas de monitoreo y control de las mejoras propuestas,

Elaboración Propia

Después de explorar la industria, y lo realizado sobre LSS, es importante entender que en la ciencia de los datos, muchos trabajos en la actualidad recorren el camino de la analítica de datos, *Big Data*, *Machine Learning*, Inteligencia artificial, y seguramente estos antecedentes pueden ser de mucha importancia para los estudios, pero es necesario verificar cuales de estos trabajos, se aproximan a nuestro objeto de estudio o tienen una relación con la optimización de procesos LSS, la logística de operaciones o directamente con la industria de servicios públicos de aseo.

Como se ha hecho con los temas anteriores, de manera tabular vamos a resumir los antecedentes y sus impactos para nuestro objeto de estudio.

Tabla 3. Antecedentes Asociados A La Ciencia De Los Datos

No.	Antecedente	Impacto
1	Minería de datos como herramienta estratégica (Flores Lagla et al., 2019)	Establece un marco claro en un lenguaje sencillo sobre los elementos básicos del descubrimiento de información para uso de toma de decisiones de valor dentro de las compañías.
2	*Data perspective of Lean Six Sigma in industry 4.0 era: A guide to improve quality* (Doga & Faruk Gurcan, 2018)	El objetivo de este estudio es proporcionar una guía que permita LSS con grandes volúmenes de datos para llegar a análisis más rápidos, confiables y satisfactorios,
3	*Towards differentiating business intelligence, big data, data analytics and knowledge discovery* (Dedić & Stanier, 2017)	Este documento de carácter descriptivo aporta claridad en los conceptos de (*Business Intelligence, Big Data, Data Analytics y Knowledge Discovery*), sus diferencias, similitudes y la relación que persiste entre ellos.
5	Aplicaciones de inteligencia artificial en procesos de cadenas de suministros: Una revisión sistemática (Icarte Ahumada, 2016)	En esta investigación se aplicaron técnicas de inteligencia artificial y algoritmos genéticos a las cadenas de suministros SC, haciendo uso de metodologías SCOR.
6	Optimizado Por Algoritmos Genéticos (Yang, 2021)	Este estudio integra algoritmos genéticos, para la optimización de procesos logísticos y determinar rutas de entrega, comparando cinco formas diferentes de planeación de rutas, un punto de partida interesante para casos de aplicación de algoritmos genéticos a problema de rutas óptimas.
7	Funcionalidades de la minería de datos (Medina Rojas & Gámez Santamaria. Cristina, 2014)	Este estudio revisa las metodologías para la minería de datos, así como los algoritmos para predicción o clúster de datos, presenta una descripción práctica que abarca desde la selección del modelo hasta el análisis e interpretación de la información obtenida.
8	Hacia un nuevo proceso de minería de datos centrado en el usuario(Aquino, 2015)	Sumado al análisis de grandes volúmenes de datos y el uso de la minería de datos, este documento articula los diseños centrados en el usuario, un elemento que es interesante verificar de cara a la integración de LSS y la ciencia de los datos
9	Introducción a la Programación de Restricciones (IBM, 2022)	Por último, esta referencia, contempla una de las herramientas de la ingeniería de software más potentes (La programación con restricciones) y la articula a la ciencia de los datos, un referente importante, si el modelo de la industria de aseo implica restricciones complejas de afrontar.

Elaboración Propia

Cómo se puede evidenciar, se cuenta con antecedentes importantes en la industria de aseo, en la optimización de procesos y en la ciencia de los datos. Pero articular estos tres enfoques para darle respuestas a los retos que tiene la industria de aseo de cara al nuevo paradigma que plantean las ciudades inteligentes y a las condiciones normativas de la industria es una motivación de fondo para justificar este proyecto.

3 ÁREA PROBLEMÁTICA Y PREGUNTA DE INVESTIGACIÓN

Ciudades Inteligentes es un concepto que cuenta con múltiples significados según el punto de vista de quien lo estudia, pero en general, dicho concepto cuenta con elementos comunes en todas las definiciones.

- La calidad de vida de quien las habita, como objetivo principal
- La producción eficiente y responsable con el medio ambiente.
- El uso de tecnología, como catalizador de los procesos.
- La importancia de los datos, en la cotidianidad.

Tal como lo manifiesta en la "Propuesta de un marco general para el despliegue de ciudades inteligentes apoyado en el desarrollo de IoT en Colombia" (Aguilar Pirachicán, 2018), La definición más completa de ciudades inteligentes la logra el UNECE (Comité Económico Europeo de las Naciones Unidas) y la UIT (Unidad Internacional de Telecomunicaciones) después de haber realizado una revisión de más de 116 definiciones.

> *"Una Ciudad Inteligente es una ciudad justa y equitativa centrada en el ciudadano que mejora continuamente su sostenibilidad y resiliencia aprovechando* ***el conocimiento*** *y los recursos disponibles, especialmente las Tecnologías de Información y Comunicación (TIC), para mejorar la calidad de vida,* ***la eficiencia de los servicios urbanos,*** *la innovación y la competitividad sin comprometer las necesidades futuras en aspectos económicos, de gobernanza, sociales y medioambientales."*

En este sentido, se evidencia el nivel de responsabilidad que las empresas de servicios públicos tienen en el marco de las ciudades inteligentes, mapeando sus dimensiones y los indicadores sobre las cuales tienen impacto directo.

• **Gestión - (Smart Governance):** gestión administrativa analítica, participación ciudadana, gobierno abierto, red de información municipal, datos abiertos y transparencia

• **Forma de Vida - (Smart Living):** seguridad ciudadana y resiliencia, mejora de la calidad de vida, salud, bienestar y accesibilidad.

• **Entorno - (Smart Environment):** reducción de gases y contaminación, reducción del impacto ambiental, gestión eficiente de los residuos

• **Movilidad - (Smart Mobility):** sistemas de planificación de rutas, concentración urbana eficiente

Estas dimensiones entre otras fueron referenciadas en el dossier central "Tecnología e innovación hacia la ciudad inteligente. Avance perspectiva y desafíos" de (Copaja Alegre & Esponda Alva, 2019), alineados completamente con el quehacer de las empresas prestadores del servicio de aseo.

Las empresas de servicios público de aseo (barrido y recolección de las ciudades), toman un papel preponderante dentro del diseño de las ciudades inteligentes y en general en la calidad de vida de los habitantes de una ciudad, por tal motivo, la regulación colombiana determina sus indicadores de productividad y eficiencia, debido a que están conectados directamente con el impacto ambiental y la calidad de vida de los ciudadanos, pero además, dicha regulación les implica una alta instrumentalización de sus operaciones, que conlleva a la generación de grandes volúmenes de información que deben ser aprovechados como conocimiento para construir ciudades resilientes.

Entendiendo el impacto que tienen sobre la calidad de vida y el medio ambiente las empresas de servicios públicos de barrido y recolección; en Colombia se regula desde 2013 con el decreto 2981 la prestación del servicio de aseo tanto "a las personas prestadoras de residuos aprovechables y no aprovechables, a los usuarios, a la Superintendencia de Servicios Públicos Domiciliarios, a la Comisión de Regulación de Agua Potable y Saneamiento Básico, a las entidades territoriales y demás entidades con funciones sobre este servicio" (MINVIVIENDA, 2013) , Así mismo, la resolución de la CRA 720 de 2015 (Comisión de Regulación de Agua Potable y Saneamiento Básico) establecen "el régimen de regulación tarifaria al que deben someterse las personas prestadoras del servicio público de aseo que atiendan

en municipios de más de 5.000 suscriptores en áreas urbanas, la metodología que deben utilizar para el cálculo de las tarifas del servicio público de aseo".(Comisión de Regulación de Agua Potable y Saneamiento Básico, 2015) . Estas regulaciones retan a las empresas de servicios públicos de aseo, a incursionar en una dinámica orientada a responder a los retos de las ciudades inteligentes.

En el siguiente cuadro, se exploran las necesidades que desde la industria de aseo se evidencian para dar cumplimiento tanto a las exigencias normativas como a los cambios de paradigma.

Tabla 4. Causas y efectos asociados al modelo de negocio

Origen	**Causa**	**Efecto - Necesidad**
Normativa	Para ser eficientes en el cumplimiento de la normatividad se requiere de una automatización de los procesos y procedimientos.	La automatización de los procesos y procedimientos operativos, requieren de una permanente retroalimentación y análisis de la información que permitan la **mejora continua de los procesos.**
Normativa	La normatividad solicita el establecimiento de sistemas de monitoreo y gestión para poder operar en ciudades de más de 5.000 habitantes.	Los modelos de operación de la industria generan grandes volúmenes de datos, en series temporales, que requieren técnicas de análisis de la información, para lograr el entendimiento de **los cambios del sistema en el tiempo** y garantizar un sistema de monitoreo efectivo.
Normativa	Lograr la calidad y la eficiencia necesaria, para garantizar el cumplimiento de las obligaciones de la CRA 720 y la rentabilidad para sus accionistas.	Las operaciones de aseo necesitan identificar las **variables significativas de su modelo** y evidenciar las relaciones existentes por medio de los datos históricos, con el fin de poder establecer la capacidad de su proceso y los puntos de mejora en términos de calidad y eficiencia.
Cambio de paradigma	**Desde la Gestión, las *Smart Governance*, requieren la** Gestión Administrativa de forma analítica, con participación ciudadana, Gobierno Abierto, que facilite las redes de información municipales, por medio de Datos Abiertos y transparencia en la información.	Hoy las empresas prestadoras de servicio requieren evidenciar una capacidad del proceso acorde a las especificaciones de calidad y una permanente preocupación por la mejora continua. Esta necesidad, vincula desde las metodologías de calidad, hasta máquinas de aprendizaje capaces de generar predicciones y prescripciones sobre las condiciones futuras del servicio.
Cambio de paradigma	**La dimensión de las formas de vida (*Smart Living*),** demandan seguridad ciudadana y resiliencia en el manejo de los servicios que mejora la calidad de vida, salud, bienestar y accesibilidad.	La resiliencia en el modelo de servicios públicos implica anteponerse a las condiciones cambiantes del servicio, por tal motivo es necesario implementar estrategias tales como análisis de datos, capaces de entrenarse a partir de la data de un área de cobertura y predecir el comportamiento del servicio en función de su impacto sobre la calidad, el bienestar y la accesibilidad del servicio.

Cambio de paradigma	**Desde la dimensión del Entorno - (*Smart enviroment*),** se requiere la reducción de gases y contaminación, la reducción del impacto ambiental y la gestión eficiente de los residuos.	Para lograr una gestión eficiente de los residuos, es necesario darles valor a los datos históricos de la organización para entender su comportamiento en el tiempo, mitigando el impacto ambiental que generan los residuos en el medio ambiente y la logística propia de la recolección de dichos residuos. Esta labor de darle valor a los datos históricos es una consecuencia de implementar estrategias del análisis de los datos, en pro de la mejora del proceso.
Cambio de paradigma	**Desde la Movilidad - (Smart Mobility), promueven la concentración urbana eficiente** por medio de sistemas de planificación de rutas, que aliviana la carga de movilidad en las ciudades.	Es necesario recordar que las operaciones de aseo cuentan con un componente logístico importante dentro de su naturaleza. En este sentido el diseño eficiente de las rutas se logra cuando el servicio es capaz de absorber los aprendizajes y convertirlos en mejoras de diseños. Para esta labor el análisis de los datos históricos, y el diseño de máquinas de aprendizaje capaces de evaluar las mejores en función de aumentar la capacidad del proceso son elementos necesarios para las nuevas empresas prestadoras de servicio de aseo en el marco de las ciudades inteligentes.

Elaboración Propia

Como se evidencia en la Tabla 4, dos grandes líneas presionan los cambios en las industrias de aseo en Colombia, en primer lugar, un marco normativo que dinamiza la industria en función de la calidad de vida de sus habitantes, y, en segundo lugar, un cambio de paradigma alrededor de las ciudades inteligentes, que promueve organizaciones más analíticas, orientadas al cuidado del medio ambiente y altamente productivas, como resultado.

En resumen, desde la dimensión de la industria, es necesario resolver la capacidad de las empresas prestadores del servicio de aseo, para analizar los datos históricos en función de diagnosticar correctamente y predecir la capacidad del proceso (en un entorno cambiante) para atender el servicio dentro de las especificaciones de calidad.

Si realizamos una revisión a los estudios más relevantes alrededor de las ciudades inteligentes a partir del 2011, son pocos los estudios que vinculan algún tipo de análisis a los servicios públicos de aseo en articulación con las ciudades inteligentes, por lo cual, se procede a identificar elementos dentro del estudio, que articulen

objetivos como eficiencia operativa, productividad, diseño de ciudades inteligentes, entre otros.

Como resultado, 11 referencias relevantes entre el 2011 y el 2019 articulan de alguna manera, los objetivos de las ciudades inteligentes y el servicio público de aseo. Dentro de esas características representativas, podemos referenciar trabajos que incluyen: la optimización de rutas (Yang, 2021), los marcos de referencia para la gobernanza (Arcidiacono et al., 2016), casos de estudio que evidencia brechas en el servicio públicos de aseo (Montalvo, 2018), estudios para las políticas de los sistemas de aseo, análisis de impacto sobre la productividad (Rivera Cerpa & Conrado Tobón, 2016), mejoras a los servicios de recolección, uso de la tecnología para gobiernos inteligentes (DNP, 2008), modelos matemáticos para rutas óptimas (Patiño Chirva et al., 2016), entre otros.

En consecuencia, existen esfuerzos diversos en términos de ciudades inteligentes y de servicios públicos, que tratan de resolver problemas de productividad, calidad, diseño de rutas o políticas administrativas alrededor de la industria de aseo. Pero no se encuentran esfuerzos consolidados en investigaciones que articulen las necesidades que plantean las ciudades inteligentes sin desconocer las exigencias normativas para las empresas de servicios públicos de aseo.

Estos retos de ciudades inteligentes y normatividad requieren conectar dos extremos de la dinámica de las operaciones de aseo. Por un lado, la optimización de sus procesos y por otro el análisis de los datos que genera la operación. El primero (los procesos) en busca de una mejora continua y un aumento de la capacidad del proceso; el segundo con el fin de generar conocimiento a partir de los datos. Cada una de estas dimensiones conlleva aplicaciones importantes para la investigación, pero además la mejora en procesos y el análisis de los datos en conjunto generan una nueva dimensión para el análisis de las empresas de servicios que atienden ciudades inteligentes.

Pese a la importancia de los servicios públicos de aseo y a los esfuerzos en regular sus operaciones, definir sus indicadores, restringir y exigir la prestación del servicio en condiciones óptimas. “Los bajos ingresos y el elevado gasto público a nivel local resultan en presupuestos reducidos para la inversión” (CEPAL, 2020).Por esa razón

optimizar los procesos y procedimientos incrementando su capacidad, es un trabajo necesario para alcanzar las eficiencias operativas que demandan nuestras ciudades.

Una de las formas para la optimización de los procesos, es la reducción de las actividades que no arrojan valor a dicho proceso, estas actividades comúnmente son llamadas desperdicios. Tal como lo manifiesta (Pérez Rave et al., 2011) a este conjunto de principios y herramientas de gestión que buscan la mejora continua a través de minimizar los desperdicios se le denomina *Lean Manufacturing*, esta metodología evidencia originalmente 7 desperdicios (Sobreproducción, Transporte, Inventarios, Esperas, Sobre Procesos, Retrabajos y Movimientos), pero en las últimas definiciones se adiciona a los desperdicios, el talento no utilizado.

Esta metodología se articuló con una filosofía de mejora continúa denominada Seis Sigma, que busca disminuir los defectos y llevarlos a 3,4 defectos por millón de oportunidades, en una estimación de la capacidad del proceso. Estas dos filosofías se complementan en un formato de mejora continúa denominado *Lean Six Sigma*, sobre el cual muchas de las industrias (en productos y servicios) trabajan para fortalecer sus estrategias de mejora continua.

De lo anterior, las primeras referencias significativas centran sus esfuerzos en la mejora de procesos para el sector logístico, muy importante para este estudio, porque como se enunció, las industrias de aseo cuentan con un gran componente logístico en su operación.

En este sentido, comprender ciertas características relacionadas con el proceso de recolección de residuos sólidos es necesaria para establecer estrategias y acciones de intervención de los procesos, optimizando recursos, ayudando a las empresas a el cumplimiento de los marcos regulatorios del país, bajo metodologías de mejora continua en base a la minería de datos de tal manera que pueda darse un mayor desarrollo de estos procesos en la ciudad, la región y el país.

Por último, se evidencia hacia el 2018 una concentración interesante de estudios, en articular la naturaleza de los datos y la mejora de procesos por medio de la ciencia de los datos y se evidencia que (Dogan & Gurcan, 2018), mapea de forma sistemática

herramientas e instrumentos de la analítica de datos, sobre la metodología *Lean Six Sigma*.

Aun así, no se localizan estudios que nos permitan articular la ciencia de los datos, la mejora de los procesos y la industria de servicios, en procura de movilizar nuevas formas de aplicación del conocimiento. Gran parte de los tiempos que deberían ser empleados para el análisis estadístico de los datos de la operación y de la mejora continua del proceso, se invierten dándole respuesta a los indicadores (claramente bien definidos) de la Comisión de regulación en su CRA 720 de 2016. Evidenciando una dificultad para buscar sistemáticamente los desperdicios de la operación (*Lean*) y en la capacidad de usar los datos como mecanismo de mejora continua (*Six Sigma*), esta situación no solo ocurre en la industria de aseo, sino que también se presenta en otras industrias como lo evidencia nuestros referentes, en la tabla de estudios relacionados con LSS descritos en la tabla 4.

Responder a las necesidades de la industria por medio de la búsqueda en la eficiencia operativa, es un camino que le permite responder a las necesidades normativas y las condiciones cambiantes de las ciudades inteligentes, pero no se debe perder de vista que la industria genera grandes volúmenes de datos y la instrumentación de sus operaciones han colocado un elemento más a esta carrera de ser una industria más competitiva, el análisis de la información.

La normativa ha llevado a las empresas de aseo a implementar sistemas de información, dispositivos móviles, monitorear el 1005 de sus flotas, sensores de temperatura, combustible, apertura y cierre de puertas, entre otro tipo de instrumentos. Colocando a las operaciones a gestionar grandes volúmenes de datos, y a identificar el valor y uso que le deben dar a la información.

En la Tabla 5. Se enuncian las necesidades de la industria desde la normatividad y desde el cambio de paradigma de las ciudades inteligentes, pero tratar de resolver esas necesidades por medio de la mejora de proceso y de la analítica de datos, evidencias nuevas oportunidades que le da punto de partida a nuestra investigación.

Tabla 5. causas y efectos asociados al manejo de los datos

Origen	Causa	Efecto - Necesidad
Los procesos	Identificar los puntos de mejora del proceso para cumplir con las condiciones actuales del servicio (EPA, 2022)	Es necesario mapear las actividades del servicio y establecer los límites de especificación que determinan la calidad de este.
Los procesos	Desde el medir es necesario establecer las variables críticas del proceso, para poder monitorear su desempeño	Realizar labores de extracción, limpieza y transformación de la información, con el fin de poder preparar la información para generar conocimiento
Los datos	La industria genera grandes volúmenes de datos desde diferentes fuentes y es necesario relacionar la información identificando patrones o comportamientos ocultos en los datos (Rudolph Raj & Seetharaman, 2013).	La minería de datos para el análisis de dichas relaciones, clasificando, clusterizando o encontrando relaciones *if-then* entre la información es una necesidad de esta industria.

Elaboración Propia

Hoy bajo la dinámica de la industria 4.0, como lo evidencia (Dogan & Gurcan, 2018), requiere de la ciencia de los datos, para permitirle encontrar nuevas relaciones, clusterizar información, clasificar los datos y entregarles a los procesos estadísticos propios de LSS información mucho más nutrida para darle valor a los análisis de capacidad del proceso.

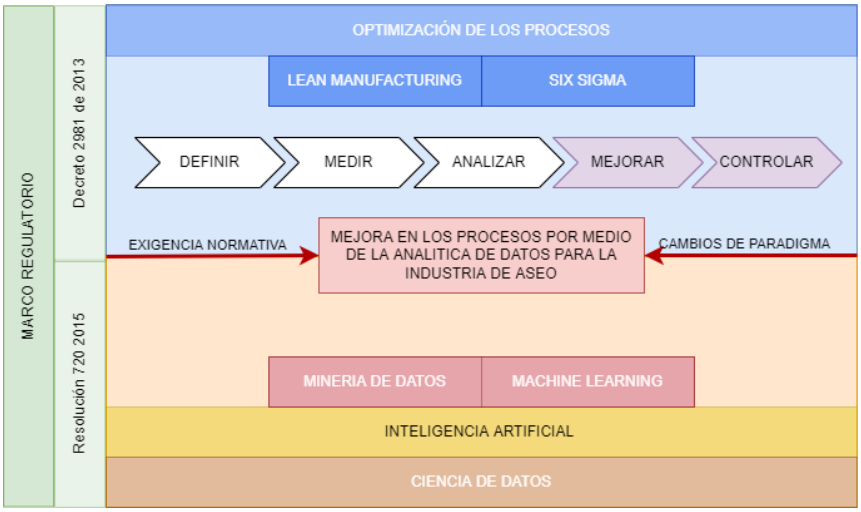
Si bien es cierto que de forma genérica, los datos transaccionales se usan para el monitoreo de las operaciones en su día a día, los análisis diagnósticos, predictivos y prescriptivos de la operación que dependen de esos grandes volúmenes de datos se desaprovechan en las operación; en la mayoría de los casos las soluciones presentan análisis descriptivos y en casos excepcionales permiten diagnosticar situaciones puntuales del cliente con ayuda de expertos que conocen la información y las operaciones (Sánchez-Muñoz et al., 2020).

No existe de fondo una ausencia de interés por esta situación, lo que sucede es que las operaciones de aseo tienen una serie de elementos cambiantes que hace que el análisis tenga un nivel de complejidad alto, con requerimientos de herramientas y recurso humano especializado que conozca los datos, la industria y sobre todo las particularidades de la operación, lo que evidencia un alto grado de dependencia de personal especializado para este tipo de procesos.

En síntesis, se integrará LSS (*Lean Six Sigma*) con la ciencia de los datos, para darle desde la investigación, una respuesta a la industria de aseo que articule estas dos ciencias en respuestas a los retos que le demanda hoy las ciudades inteligentes y la normativa vigente en Colombia.

Por tanto, en la presente investigación se estudia el desarrollo de un proceso de mejora continua bajo el modelo de *Lean Six Sigm*a (LSS), integrando la metodología de desarrollo DMAIC de mejora de proceso (Definir, Medir, Analizar, Mejorar, Controlar) (Garza Ríos et al., 2016) y las técnicas de análisis de datos, lo cual permite articular los procesos de mejora continua a través del desarrollo de máquinas de aprendizaje capaces de predecir el comportamiento del sistema a partir de las variables sugeridas por el modelo, esto con el fin de optimizar el proceso de recolección de residuos sólidos urbanos en este caso, para la ciudad de Manizales como se muestra en la Figura 1.

Figura 1. Área Problemática



Elaboración propia

En este contexto, surge la siguiente pregunta de investigación:

> *¿Qué acciones de mejora en la capacidad del proceso para una operación de recolección de residuos sólidos, pueden determinarse integrando Lean Six Sigma y ciencia de datos?*

4 JUSTIFICACIÓN

Los procesos de recolección de RSU se han constituido como una de las problemáticas que deben ser atendidas por los países de forma prioritaria debido a la industrialización, el crecimiento de la población urbana y el consumo, para lo cual se requiere la adopción de estrategias eficientes que permitan su aprovechamiento y disposición adecuada. Por tanto, las administraciones locales, regionales y nacionales deben tomar decisiones de forma que se realice una gestión correcta de estos residuos además que tengan presente las particularidades de su región. Dentro de estas particularidades se encuentra, entre otros elementos, la caracterización de la población, la cantidad de RSU generados, el clima, las mejores técnicas y tecnologías disponibles, las tendencias y políticas nacionales, la disponibilidad de recursos económicos, los planes de ordenamiento territorial, la legislación vigente y las características del servicio público de limpieza (Solano, 2021)

Desde la política pública nacional diseñada para el manejo de los residuos sólidos se incluye la necesidad de generar inversiones e instrumentos económicos con el fin de lograr un crecimiento sostenible y un desarrollo económico (DNP, 2016). Si bien existen muchos factores, el desarrollo de ciudades inteligentes es un aspecto emergente del desarrollo económico y una forma de satisfacer las necesidades de las personas. De acuerdo con lo anterior, se hace evidente incorporar y caracterizar procesos en este caso el de RRSU en cada región, ya que cada uno es característico por sus dinámicas diferenciadoras además de ser uno de los procesos con mayor potencialidad de crecimiento en adopción de nuevas tecnologías(Sánchez-Muñoz et al., 2020), además que cuentan con un nivel de impacto estratégico al momento de entablar los diálogos de ciudades inteligentes. En este sentido, se convierten en un foco importante para la gestión de nuevos proyectos de investigación y desarrollo tecnológico que generan el progreso de centros urbanos en los que la concentración demográfica y económica generan desafíos a los cuales se debe hacer frente.

Es importante entender que la industria de aseo en Colombia es uno de los servicios públicos domiciliarios con mayor potencialidad de crecimiento en adopción de nuevas tecnologías, y cuentan con un nivel de impacto estratégico al momento de entablar

diálogos de ciudades inteligentes. Esta industria moviliza en principio 3 Indicadores de ODS.

- ODS 6 Agua limpia y saneamiento Básico, minimizando el impacto sobre los humedales producto de la mala disposición de los residuos sólidos

- ODS 11 Ciudades y comunidades sostenibles, este es uno de los ODS a los cuales la industria de aseo impacta de una forma directa, el 55% de la población mundial vive en las ciudades.

- ODS 12. Producción y Consumo responsable, son las empresas de aseo un sensor de las ciudades en relación con la cantidad de residuos que logramos entrar en economías circulares.

Como lo manifestó en su informe de gestión La Superintendencia de servicios públicos desde el 2019 (Alejandro & Urrea, 2019)

Por esto la industria de aseo y recolección de residuos en las ciudades, se convierten en un foco importante de investigación, hacerlas más competitivas, más eficientes y mejorar las capacidades de sus procesos nos permiten tener ciudades mucho más sostenibles y responsables con el medio ambiente.

Pero no solo el impacto a los objetivos de desarrollo sostenible hace parte de los intereses de las empresas de servicios públicos sobre este trabajo, para ellas, el análisis de la información y la identificación de relevancias y correlación de las variables pueden significar el incremento en la capacidad del proceso y como consecuencia, contar con una operación controladas en sus variables críticas de calidad.

Llegar a este nivel de madurez se logra cuando las empresas que administran y gestionan la información, están en la capacidad de ofrecer análisis detallados fundamentados en ciencia de datos, esta investigación arroja un punto de partida importante para determinar patrones de correlación entre las variables, pero además integra técnicas de minería de datos, muy útil para el despliegue de estas técnicas en industrias similares.

En la medida que las ciudades inteligentes comienzan su evolución, el volumen de datos se incrementa exponencialmente; Es importante contar con investigaciones que apliquen la ciencia de los datos para simplificar su análisis y procesamiento. Esta investigación explora las relaciones entre variables y la relevancia que tienen al momento de calcular la capacidad del proceso, integrando metodologías de capacidad del proceso como LSS, con técnicas de minería de datos. Si bien la utilidad de la investigación está percibida por los concesionarios de aseo, las empresas que gestionan sus sistemas y la sociedad como impacto a sus objetivos de sostenibilidad, los investigadores también podrán encontrar en esta investigación elementos novedosos.

El primero es la integración de una metodología tradicional de medición de capacidad de procesos, con la Ciencia de los datos, con el objetivo de poder estimar de forma correcta las relaciones de dependencia entre las variables del modelo de la industria de aseo, además la posibilidad de que un modelo computacional, permite explorar la relevancia de dichas relaciones por medio del impacto que generan en la capacidad de los procesos.

En el CONPES 3530 de 2008, lineamientos y estrategias para fortalecer el servicio público de aseo en el marco de la gestión integral de residuos sólidos, se estableció como objetivos, “establecer los mecanismos para evaluar y ajustar la calidad, viabilidad y seguimiento de los esquemas de planeación integral del servicio público de aseo, mejorando la articulación de las inversiones”, desde ese momento hasta la fecha garantizar la eficiencia y la eficacia de los servicios de aseo ha sido una prioridad (Oliver, 2008) . Esto establece un punto de partida para los análisis predictivos y prescriptivos que se pueden afrontar en nuevas investigaciones incluyendo técnicas de Inteligencia Artificial y Machine Learning para lograr procesos de aprendizaje reforzado.

Actualmente la industria, desde su dimensión operacional, ha centrado sus investigaciones y esfuerzos en determinar la eficiencia y optimización de las rutas, pero no se encuentra estudios documentados donde se mida la capacidad de los procesos de recolección de residuos, este es un punto de partida para la industria que

nos acerca más al concepto de contar con procesos controlados, sujetos de optimizaciones y controles automatizados, de cara a impactar la productividad de la operación.

Hoy se cuenta con acceso a la información detallada de los procesos de recolección de un territorio, con más de 5 años de historia, un elemento privilegiado para poder determinar el valor de los datos para la industria, la academia y la sociedad.

5 REFERENTE TEÓRICO

5.1 SISTEMAS DE INFORMACIÓN GEOGRÁFICO

Los SIG son un conjunto de herramientas destinadas al procesamiento de datos geográficos. Son responsables de integrar varios aspectos de los datos geográficos del mundo real, permitiendo organizar, recopilar, manipular, analizar y modelar dichos datos. Se constituyen como una herramienta multipropósito ya que pueden trabajar con todo tipo de datos geográficos del mundo real, incluidos los relacionados con la tierra tales como datos topográficos y geológicos, así como también datos estadísticos y epidemiológicos relacionados con individuos, por ejemplo la propagación de enfermedades en una región de tierra (M.Howari & Ghrefat, 2021).

Primero, un SIG se construye modelando datos geográficos del mundo real, generalmente con la ayuda de un sistema de coordenadas y con datos recopilados de instrumentos de detección remota y un sistema de posicionamiento geográfico. Los datos espaciales resultantes incluyen información relacionada con las relaciones geográficas, como la ubicación, la forma, el tamaño y la orientación de un punto de interés geográfico. Luego, puede proceder a analizar dichos datos espaciales compuestos o en capas, y al manipular dichos datos, además puede generar todo tipo de patrones geográficos e información relevante para los tomadores de decisiones (M.Howari & Ghrefat, 2021).

De aquí, la versatilidad con que cuentan los SIG, dependiente del tipo de dato, ya que en el caso de los datos vectoriales nos permite de la segmentación de los espacios bidimensionales a través de polígonos representada como información de segmentación gráfica (país, ciudad, reserva natural, entre otros), añadir información a través de capas, estas capas nos permiten una mejor estructuración y visualización de nuestro conjunto de características. Por otro lado, para los modelos de datos ráster, la información se contiene en matrices de píxeles que pueden ser representadas a través de la intensidad de colores, para representar una característica numérica o diferenciación de los espacios.

Actualmente, en el mercado se encuentra Geoaseo una herramienta tecnológica que logra optimizar el ejercicio de las empresas de aseo en operaciones como gestión de rutas, recolección, barrido, atención de aforos, servicios especiales y seguimiento

vehicular entre otras variables que se presentan en el proceso de aseo en las ciudades y municipios del país (SIGMA INGENIERIA S.A, n.d.)

Es un sistema de información basado en SIG y con entornos web y móvil, posibilita la gestión de los procesos operativos, capturando y procesando datos geográficos y alfanuméricos generando información, alineados a los indicadores de gestión de cada empresa. GEOASEO concentra sus soluciones en mejorar la calidad, eficacia, productividad y rentabilidad de su organización

Figura 2 Software Geoaseo

Elaboración Propia

Actualmente, GEOASEO cuenta con diferentes módulos configurados, como se describe en la Figura 3. En el Anexo 1 – Se encuentra la descripción del producto Geoaseo.

• Recolección

•Seguimiento vehicular

•Atención móvil

•Mantenimiento

•Barrido

•Inteligencia de Negocios

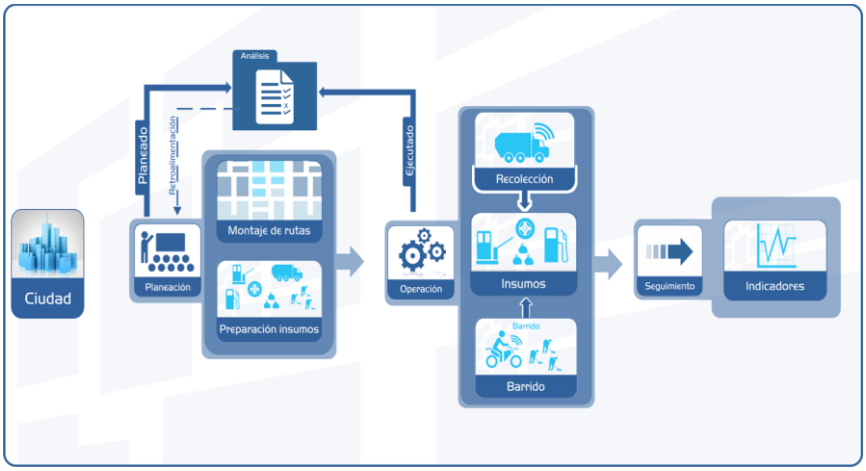
•Servicios complementarios

•Supervisión

•Respel

• Aforos

Figura 3 Módulos Software Geoaseo



SIGMA INGENIERIA

El sistema anteriormente descrito, genera un gran volumen de datos, de los cuales gran utilidad para este proyecto de investigación se da en el módulo de recolección. A continuación, se presentan las dos sabanas de datos, objeto de estudio para el desarrollo de esta investigación

- **Hoja De Ruta Planeada:** es una estructura de información, que le permite al profesional encargado del diseño de la operación, explicar las condiciones de tipo de ruta, hora de inicio, hora de finalización, kilómetros recorridos, tiempos de la operación, cantidad de combustibles planeados, número de operarios, entre otros y que evidencia para cada una de las zonas de la ciudad y de las micro rutas las condiciones ideales de operación. Por lo regular este instrumento es generado por un especialista en SIG acompañado de toda la estructura cartográfica de la operación.

- **Hojas De Rutas Ejecutadas:** complementan las variables de ejecución de la operación que provienen de instrumentación como GPS y tablet dispuesta para cada uno de los conductores, en esta hoja de recolección ejecutada se almacena la información de km recorridos, combustible, hora de inicio, hora de fin, ciclos de la operación, así como la información relacionada con las toneladas recogidas en cada uno de los viajes reales. Este instrumento sube automáticamente en el sistema Geoaseo plan de datos desde los dispositivos móviles o GPS, capturando cada 50 mts o cada interacción del operario del vehículo, para finalmente totalizar la operación y llenar las variables de forma automática en la recolección ejecutada.

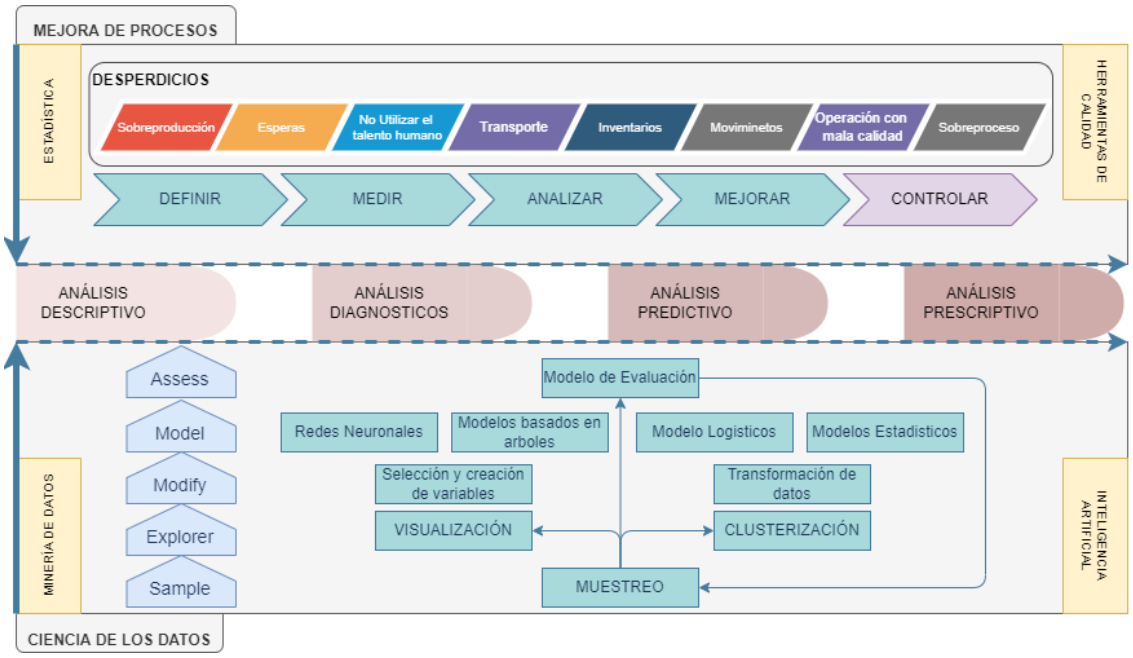
El proceso de recolección y transporte es la actividad principal del modelo de negocio de aseo. Es la actividad principal del servicio pues participa con un 40% en la estructuración de costos. Se basa en la recolección de residuos en el domicilio del generador y en el transporte a los sitios de disposición final. Por regulación gubernamental la recolección debe realizarse en vehículos compactadores, pero aún es frecuente utilizar volquetas descubiertas, principalmente en municipios de categoría 4, 5 y 6. La unidad de medida para esta actividad es la tonelada recolectada y transportada.

Lo anterior pone en manifiesto la importancia de analizar los datos para dar respuesta a los retos que hoy se plantean para las empresas de servicios públicos de aseo, requiere de la integración de dos grandes líneas de conocimiento, la mejora de procesos y la ciencia de los datos. Para la mejora de procesos la línea de trabajo seleccionada es LSS que nos permite definir, medir, analizar y plantear las opciones de mejora para los objetivos de calidad del cliente, identificando los desperdicios más significativos en el proceso (sobreproducción, esperas, No uso del talento humano, transporte, inventarios, movimientos, operación con mala calidad, sobre procesos), todo este esfuerzo apoyado por herramientas estadísticas y de calidad. Queda por fuera del alcance de este estudio las actividades asociadas al control.

Por su parte, desde la ciencia de los datos, se dé un conjunto de datos suministrados por la industria, visualizándose y agrupándose, para transformarlos a través de la selección o creación de nuevas variables, modelando el sistema por medio de las

técnicas más adecuadas (redes neuronales, regresiones logísticas, árboles de decisión, modelos estadísticos).

Figura 4 Diagrama Del Marco Teórico



Elaboración Propia

Todo esto con el objetivo de darle respuesta a partir de la información a los análisis descriptivos, diagnósticos, predictivos o prescriptivos, del modelo de aseo que requiere la industria actual. En la parte superior de la Figura 4, tenemos a *Lean Six Sigma*, qué es la combinación de dos metodologías de mejora de procesos, cada uno de ellos busca la maximización de la productividad de forma independiente, pero al combinarlas buscan una reducción de los costos y un aumento de la capacidad del proceso.

5.2 LEAN MANUFACTURING

Con el fin de clarificar los conceptos de *LSS se* adoptarán las definiciones que se encuentran en la Tabla 6.

Tabla 6. Definiciones Básicas

No.	Concepto	Definición
1	*Lean Manufacturing*	"*Lean Manufacturing* es una filosofía de trabajo, basada en las personas, que define la forma de mejora y optimización de un sistema de producción focalizándose en identificar y eliminar todo tipo de "desperdicios", definidos éstos como aquellos procesos o actividades que usan más recursos de los estrictamente necesarios."
2	*Six Sigma* (Bento da Silva et al., 2019)	Significa "seis desviaciones estándar". La visión *six sigma* busca llevar nuestros procesos a una tasa de error de 3,4 defectos por millón de oportunidades
3	Capacidad del proceso Cp y Cpk	Es el cociente entre el rango de tolerancia admitida para el proceso y su propia capacidad.

Elaboración Propia

Recorre toda la cadena de valor del producto o servicio, identificando ocho tipos de desperdicios que no arrojan valor para el cliente, pero como resultado inyectan tiempos y costos de producción innecesarios para lograr los objetivos. Los desperdicios se clasifican en sobreproducción, esperas, no utilización correcta del talento humano, transporte, inventarios, movimientos, operaciones con mala calidad, sobre procesos (Aouag & Mohyiddine, 2023).

En un enfoque tradicional, estos desperdicios son identificados y categorizados, por medio de un grupo de expertos, en la industria 4.0 este tipo de desperdicio debe ser localizado desde el análisis de los datos, para lograr hacerlo de manera sistemática y recurrente.

Six Sigma, por su parte busca la eliminación sistemática de los fallos en el producto o servicio, centra sus esfuerzos en encontrar y reducir la variabilidad del proceso (Tampubolon & Purba, 2021), para lograr esta meta es necesario, apoyarse en herramientas de calidad y de estadística en cada una de las etapas:

- La primera parte para la búsqueda de los fallos es, **Definir** el problema por medio de los criterios de calidad del cliente, para esta etapa se hace uso de herramientas de calidad como estatus del proyecto (Project Charter), matriz

de la voz del cliente (VOC), despliegue de función de la calidad (QFD), análisis de riesgo, matriz de viabilidad del proyecto y mapa de procesos (SIPOC, VSM).

Las herramientas de Big Data también son importantes en las etapas de definición porque permiten técnicas como minería de texto, minería de videos y procesos de descubrimiento de información.

- Después se procede a **Medir** por medio de la recolección correcta de los datos, asegurando la exactitud de las mediciones, en esta etapa las herramientas de calidad como plan de recolección de datos y las herramientas estadísticas estudios R & R, estadística básica, capacidad del proceso, análisis gráficos, diagramas de Pareto y gráficos de control, toma una relevancia importante.
- Todos los elementos anteriores suministran una entrada esencial para la etapa de **Analizar** y verificar las causas potenciales de falla, por medio de herramientas estadísticas como pruebas de hipótesis, análisis descriptivo, análisis de regresión, diseño de experimentos (DOE), o cuantificación de oportunidades, que nos permiten identificar la causa raíz de los problemas, todo con el objetivo de proponer acciones de **Mejora**

La etapa de análisis es quizás una de las etapas que más apoyo recibe de la ciencia de los datos, desde la minería de datos, las reglas de asociación, la clusterización y la clasificación son técnicas de alto valor; pero no solo la minería de datos, desde el Big Data Las máquinas de aprendizaje, los árboles de decisión, la minería de texto y de video, así como la inteligencia artificial también tienen cabida en esta etapa del proceso.

- Esta penúltima etapa, la **Mejora** permite llegar a los objetivos de calidad definidos inicialmente, vinculado herramientas como Teoría de restricciones (TOC), 5S, Reducción del tiempo de puesta a punto, sistemas genéricos “pull”, poka yoke, entre otros. Pero desde la ciencia de los datos y específicamente desde Big Data las máquinas de aprendizaje y la inteligencia artificial hacen su aporte de manera contundente.

Todas estas acciones no serían sostenibles sin un proceso de **Control** que permita realizar un seguimiento permanente a las acciones de mejora y garantizar la implementación de las mejoras.

La ciencia de los datos por su parte hace uso de la metodología SEMMA (Sample, Explorer, Modify, Model, Assess) vinculando todas las técnicas enunciadas anteriormente para generar un modelo que responda a los análisis descriptivos, diagnósticos, predictivo y prescriptivos del sistema a estudiar (Palacios et al., 2017).

5.3 CIENCIA DE DATOS

La operación de aseo hoy desde los SIG, trae información desde la planeación del proceso y la ejecución real , el análisis exploratorio de los datos genera un panorama general del proceso, articulado las técnicas de análisis de datos con los procesos de mejora continua como LSS en este caso definiendo la capacidad del proceso desde la información real, además de involucrar diferentes técnicas basadas en inteligencia artificial para implementar acciones de mejora en las operaciones, a continuación se presentan algunos procedimientos importantes aplicados en este estudio:

5.3.1 Estadística De La operación: Rutas Planeadas Y Ejecutadas

Previo a entrar a manipular y trabajar con la base de datos con fines científicos, se hace uso de la Estadística Descriptiva, la cual se compone de un grupo de métodos que permiten describir dichos datos de manera reducida y ordenada, esto con el fin de encontrar características representativas. Dado que de acuerdo con el tipo de datos que se tiene se tendrán distintos métodos y gráficos adecuados para el análisis de dicho conjunto, se deberá describir la clasificación de variables, así como también cuáles medidas son las medidas que le involucran: tendencia central, dispersión, desviación estándar y forma (Calvo Mazuera, 2020).

Cabe destacar que cada característica de una base de datos (población) es una variable, y a cada muestra de la población se le denomina registro.

Tipo De Variables

- Variables Cualitativas: Están son el tipo de variables que definen cualidades de los registros, a su vez se subdividen en Ordinales y Nominales, de esta manera para la primera existe una relación de orden dentro de las categorías, mientras que para la segunda se asigna un nombre el cual no está delimitado a un criterio de orden.
- Variables Cuantitativas: Estos son los atributos que son medibles o cuantificables numéricamente, y se subdividen en Discretas o Continuas, la primera de estas para números contable de valores como número de personas, llamadas, registros; mientras que para la segunda cuando se puede asumir un valor cualquiera con nivel de precisión dado, como el número de kilómetros recorridos, tiempo total en horas, entre otras.

Agrupación De Datos Y Distribución De Frecuencias - Gráficos

Dentro del conjunto de datos a menudo podemos encontrar características generales agrupando los datos en un determinado número de clases, intervalos o categorías. De aquí, la distribución de frecuencias sustenta el describir numérica y gráficamente dichas agrupaciones. Así pues, dependiendo de las variables se tiene:

- **Gráfico De Barras:** Consiste en la construcción de un gráfico que muestra el resultado de contabilizar la cantidad de muestras que existen para las categorías específicas de una variable. Este gráfico es ejecutable para variables cualitativas o cuantitativas discretas.
- **Gráfico Circular:** Gráfico que divide la superficie del círculo en proporción a la cantidad de muestras para una categoría dada. Este gráfico es ejecutable para variables cualitativas o cuantitativas discretas.
- **Histograma:** Dicho gráfico se construye mediante columnas o rectángulos unidos. Así pues, se divide el rango total de una variable continua en determinados intervalos, y sobre cada intervalo o clase, se levanta un rectángulo que tiene como base la respectiva amplitud del intervalo y como altura la cantidad de muestra existentes que se encuentran en dicho intervalo para dicha variable.

Medidas De Tendencia Central

- Media de datos no agrupados: es la medida de tendencia más representativa, la idea de media o promedio formaliza el concepto intuitivo de punto de equilibrio o centro de gravedad. Se define matemáticamente como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad (1)$$

- Mediana: es la medida que indica la posición central de los datos
- Moda: indica el valor de la variable con mayor frecuencia absoluta, se puede utilizar tanto para datos numéricos o categóricos.

Medidas De Dispersión

Son aquellas que muestran la variabilidad de una distribución, así pues, cuanto mayor sea el valor, mayor variabilidad, cuanto menor sea dicha cantidad, más homogénea. Algunas de ellas son Rango, Desviación, Varianza.

- Varianza Muestral: es el promedio del cuadrado de las desviaciones respecto a la media, generando una medida única de todas las desviaciones del conjunto de datos. Matemáticamente:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \quad (2)$$

- Desviación Estándar: es la raíz cuadrada positiva de la varianza, procede a la misma conceptualización de la varianza, asimismo decimos que la desviación estándar es pequeña si los valores se concentran alrededor de la media. Matemáticamente:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \quad (3)$$

- Coeficiente De Variación: es el cociente de la desviación estándar y la media. Es una medida de dispersión comparativa que mide diferencias en variabilidad cuando en grupos de datos diferentes poseen la misma media.

$$CV = \left(\frac{\sigma}{\bar{x}}\right) * 100\% \quad (4)$$

Medidas De Forma

Estas contienen un marco de referencia para calcular matemática y gráficamente la forma de una distribución dada. Entre ellas encontramos los coeficientes de asimetría y curtosis.

- Curtosis: denota la forma de la curvatura en la distribución que forma las frecuencias. Matemáticamente:

$$Kur = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3 \quad (5)$$

- Coeficiente De Asimetría: describe la manera como los datos tienden a reunirse de acuerdo con la frecuencia. De forma analítica se puede calcular por medio de varios coeficientes, tales como Coeficiente de Asimetría de Fisher, Coeficiente de Asimetría de Pearson y Coeficiente de Asimetría de Bowley, siendo el primero de ellos el más utilizado, matemáticamente:

$$CA_f = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\sigma^3} \quad (6)$$

5.3.2 Capacidad Del Proceso

El análisis de capacidad de un proceso se enmarca en el control estadístico de la calidad y se define como la resolución en la cantidad de la variación natural de un proceso para una característica de calidad dada, esto permitirá saber en qué medida tal característica de calidad es satisfactoria (cumple especificaciones). En otras palabras, son mediciones especializadas que sirven para evaluar de manera práctica la habilidad de los procesos para cumplir con las especificaciones (Gutiérrez Pulido & de la Vara Salazar, 2013).

Dentro de los índices que nos ayudan al análisis de capacidad de un proceso encontramos:

- Índice de capacidad potencial o Cp: es el índice de capacidad potencial del proceso es el resultado matemático de encontrar dividir el ancho de las especificaciones (variación tolerada, esta es la característica de calidad) entre la amplitud de la variación real del proceso, la cual se cuantifica como:

$$C_p = \frac{LSE - LIE}{6\sigma} \quad (7)$$

De donde se representa la desviación estándar del proceso, mientras que LSE y LIE son las especificaciones superior e inferior para la característica de calidad. Otra interpretación que podemos dar a este índice de capacidad potencial es el de la proporción entre el rango en las especificaciones (o variación tolerada) y el de la variación real (recordando las propiedades de la distribución normal, en donde se afirma que entre se encuentran el 99.73% de los valores de una variable con distribución normal. Incluso si no hay normalidad) (Gutiérrez Pulido & de la Vara Salazar, 2013).

Tabla 7. Índices De Capacidad Del Proceso

Valor de Índice Cp	**Clase o Categoría del Proceso**	**Decisión (Si el proceso está centrado)**
$C_p > 2$	Clase Mundial	Tiene calidad Seis Sigma.
$C_p > 1.33$	1	Adecuado.
$1 < C_p < 1.33$	2	Parcialmente adecuado, requiere de un control estricto.
$0.67 < C_p < 1$	3	No es adecuado para el trabajo. Es necesario un análisis del proceso. Requiere de modificaciones serias para alcanzar una calidad satisfactoria.
$C_p < 0.67$	4	No es adecuado para el trabajo. Requiere de modificaciones muy serias.

(Gutiérrez Pulido & de la Vara Salazar, 2009)

Índice de capacidad real o Cpk: se define como el indicador de la capacidad real de un proceso que se puede ver como un ajuste del índice Cp para tomar en cuenta el centrado

del proceso. Se conoce como índice de capacidad real del proceso y es considerado una versión corregida del Cp (Gutiérrez Pulido & de la Vara Salazar, 2013). A pesar de que existen muchas formas equivalentes para calcular dicho índice, una de las más comunes es la siguiente:

$$C_{pk} = min[\frac{\mu - LI}{3\sigma}, \frac{LS - \mu}{3\sigma}] \quad (8)$$

Índice de desempeño potencial o Pp.

Índice de desempeño real o Ppk.

Para objeto de este estudio solo se definirán matemáticamente estas dos primeras, ya que serán estas las que nos permitirán evaluar el proceso.

5.3.2 Análisis De Componentes Principales, Análisis Multivariados, Normalización

- **Análisis De Componentes Principales**

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de análisis multivariado que transforma linealmente un conjunto de variables correlacionadas en un conjunto de variables no correlacionadas. Buscando que un conjunto con un número menor de variables no correlacionadas puede explicar una parte sustancial de las informaciones del conjunto original, a esto último se le conoce como variabilidad de los datos (Cheng, 2022). Los objetivos más relevantes del Análisis de Componentes Principales son:

- Reducción de la dimensionalidad de los datos.
- Obtención de combinaciones lineales que tienen interpretación de las variables originales.
- Descripción y entendimiento de la estructura de correlación de las variables originales.

Matemáticamente, este método se sustenta en el teorema de descomposición espectral el cual transforma el conjunto de variables originales en un otro conjunto de variables,

para lo cual, cada una de estas nuevas variables resultan ser combinaciones lineales de las primeras, cabe resaltar que esto no hace la variabilidad presente en el conjunto inicial cambie. Las componentes principales dependen únicamente de la matriz de covarianza Σ o de la matriz de correlación ρ. La matriz Σ se puede emplear cuando todas las variables tienen unidades de medida similares y también cuando se estime conveniente destacar cada una de las variables en función de su grado de variabilidad y la matriz ρ se usa para dar la misma importancia a cada una de las variables; esto puede ser conveniente cuando se quiera considerar que todas las variables sean igualmente relevantes.

Este método ha sido aplicado en diferentes campos como lo son: psicología, medicina, meteorología, geografía, taxonomía, biología, finanzas, agricultura, ecológica y arquitectura. Demostrando especial relevancia en genética, debido a la enorme cantidad de información que se maneja.

Cabe resaltar que esta técnica se utilizará en el presente trabajo como una herramienta para analizar las correlaciones existentes entre el conjunto de datos, para ello, con principal enfoque en el análisis de la mayor variabilidad de datos con el menor número de variables, si hay lugar a ello.

- **Análisis Multivariados**

Los análisis multivariados hacen parte de la estadística multivariante, esta se usa para abordar las situaciones en las que se realizan múltiples mediciones en cada unidad experimental y las relaciones entre estas mediciones y sus estructuras son importantes (Grech & Calleja, 2018). Existen diferentes tipos de análisis para lo cual podemos clasificarlos en 3 grupos:

    - Métodos de Dependencia.
    - Métodos de Interdependencia.
    - Métodos Estructurales.

- **Pruebas De Normalidad**

Las pruebas de Normalidad son pruebas estadísticas para evaluar si una distribución de datos cumple o no, con las características propias de una distribución normal (también conocida como distribución gaussiana) (Greener, 2020), algunas características que se destacan de dicha distribución son:

- La media es central: La línea media que divide la función en dos partes iguales y simétricas.
- Unimodal: La mayoría de los datos se encuentran concentrados en el centro.
- Asintótica: La función para sus colas izquierda y derecha es asintótica al eje de las abscisas.

Existen diversas pruebas de este estilo como las pruebas de hipótesis, la prueba de curtosis y la asimetría. Para las primeras nos encontramos las de: Kolmogorov-Smirnov, Kolmogorov-Smirnov con corrección de Lilliefors y Shapiro Wilk, cada una con sus respectivas ventajas y desventajas.

Normalización Puntuación Z

La técnica de normalización en puntuación Z es una transformación numérica que realiza un escalado del conjunto de datos, esto con el fin de evitar sesgos en el caso que el rango de los datos numéricos sea muy grande, en dicho caso los pesos de los valores más grandes son mayor de aquellos números cercanos al rango inferior. Entre otras cosas esta transformación nos permite comparar distribuciones de frecuencia, ya que entre sus características encontramos que la media del resultado de la transformación por puntuación z se encuentra en 0 y la desviación estándar original se caracterizará en dicha transformación por el valor de +1 y -1 (Scikit Learn, 2022b). Matemáticamente:

$$z = \frac{x-\mu}{\sigma} \tag{9}$$

5.3.3 Modelos De Aprendizaje Automático

- ***Support Vector Regression***

Para poder hablar de Support Vector Regression (SVR), debemos introducir un concepto aún más intuitivo, este es el de Support Vector Machine (SVM), el cual es un algoritmo de aprendizaje automático supervisado, ampliamente utilizado tanto para problemas de clasificación como de regresión. De manera general, un SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación, definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte (Belyadi & Haghighat, 2021).

Este concepto se extiende como método de regresión, de manera tal que se mantienen todas las características, utilizando los mismos principios, pero exceptuando algunas diferencias estableciendo un margen de tolerancia (épsilon) de aproximación al SVM que ya habría pedido el problema. Así pues, la idea principal es siempre la misma: minimizar el error individualizando el hiperplano que maximiza el margen, teniendo en cuenta que se tolera parte del error (Sayad, 2022). De tal manera, lo que se busca matemáticamente es, para SVR lineal:

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot \langle x_i, x \rangle + b \tag{10}$$

Para SVR no lineal:

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b \tag{11}$$

- ***Stochastic Gradient Descent Regression***

El Descenso de Gradiente Estocástico (SGD) es un método iterativo para optimizar una función objetivo con propiedades como diferenciable o subdiferenciable. Se podría considerar como una aproximación estocástica del método de optimización del descenso del gradiente, ya que reemplaza el gradiente real (calculado a partir de todo el conjunto de datos) por una estimación de este (calculado a partir de un subconjunto de datos seleccionado al azar). Especialmente en problemas de optimización de alta dimensión,

esto reduce la carga computacional muy alta, logrando iteraciones más rápidas en el intercambio por una tasa de convergencia más baja (Sra et al., 2011). Así pues, matemáticamente una función a optimizar:

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w) \tag{12}$$

En donde por medio de la optimización del descenso del gradiente, se representa:

$$\omega \coloneqq \omega - \eta\nabla Q(\omega) = \omega - \frac{\eta}{n}\sum_{i=1}^{n} \nabla Q_i(\omega) \tag{13}$$

Por el método el descenso de gradiente estocástico se tendría:

$$\omega \coloneqq \omega - \eta\nabla Q_i(\omega) \tag{14}$$

- ***Bayesian Ridge Regression***

La regresión bayesiana, también conocida como regresión lineal bayesiana, es un enfoque que utiliza la inferencia bayesiana en la construcción de modelos regresión lineal, mecanismo útil cuando se tienen datos insuficientes o datos mal distribuidos mediante la formulación de una regresión lineal utilizando distribuciones de probabilidad en lugar de estimaciones puntuales (Bedoui & Lazar, 2020). Se supone que la salida o respuesta 'y' se extrae de una distribución de probabilidad en lugar de estimarse como un valor único, esto permite mayor flexibilidad al momento de realizar estimaciones (Pedregosa et al., 2011). Matemáticamente, para obtener un modelo probabilístico, se parte de la idea que la salida ‘y’ es asumida como una distribución Gaussiana alrededor de ‘Xw’, así:

$$p(y|X, w, \alpha) = N(y|Xw, \alpha) \tag{15}$$

De esta, la forma general de la regresión bayesiana el coeficiente ‘w’ está dado por lo que se denomina un Gaussiano esférico, de esta manera:

$$p(w|\lambda) = N\left(w\middle|0, \lambda^{-1}I_p\right) \tag{16}$$

- ***LassoLars***

Lasso es la técnica de operador de contracción mejor estudiada y más básica. LASSO reduce los coeficientes (parámetros) que se correlacionan con cero o cerca de cero, dando como resultados estimadores con variantes más pequeñas y un modelo final más representativo. El método Lasso se hizo conocido después del algoritmo LAR (Regresión de ángulo mínimo) en 2004. Las rutas de solución de LAR son lineales por partes y, por lo tanto, se pueden calcular de manera muy eficiente. LARS es un algoritmo eficiente para estimar parámetros LASSO computacionales. El método LASSO puede reducir el coeficiente del método de mínimos cuadrados ordinarios a cero para que pueda seleccionar la variable fija (Adelheid Januaviani et al., 2019), (Scikit Learn, 2022a) .

El Lasso es un modelo lineal que estima coeficientes dispersos. Es útil en algunos contextos debido a su tendencia a preferir soluciones con menos coeficientes distintos de cero, reduciendo efectivamente el número de características de las que depende la solución dada (Pedregosa et al., 2011)]. Matemáticamente consiste en un modelo lineal con un término de regularización añadido. La función objetivo a minimizar es:

$$min_w \frac{1}{2n_{samples}} ||X\omega - y||_2^2 + \alpha||\omega||_1 \quad (17)$$

Por otro lado, el algoritmo de ángulo mínimo (LARS) se utiliza para ajustar modelos de regresión lineal a datos de alta dimensión. El cual, en cada paso, encuentra la característica más correlacionada con el objetivo. Cuando hay varias características que tienen la misma correlación, en lugar de continuar a lo largo de la misma característica, avanza en una dirección equiangular entre las características (Pedregosa et al., 2011).

- ***Automatic Relevance Determination Regression***

Este es un caso de especial de ajuste de la 'Bayesian Ridge Regression', que conduce a coeficiente 'w' dispersos, ya que desde el modelo se plante este coeficiente de manera distinta, dejando a un lado el supuesto de distribución Gaussiana esférica, en vez de esto el 'w' se asume como distribución Gaussiana elíptica paralela al eje. Esto significa que cada coeficiente 'wi' es dibujado por una distribución gaussiana, centrado en cero con una precisión 'λi', así:

$$p(w|\lambda) = N(w|0, A^{-1}) \tag{18}$$

De esta manera, la estimación se realiza mediante un procedimiento iterativo (Pedregosa et al., 2011).

- ***Passive Aggressive Regression***

Los algoritmos pasivo-agresivos se utilizan generalmente para el aprendizaje a gran escala. En los algoritmos de aprendizaje automático en línea, los datos de entrada vienen en orden secuencial y el modelo de aprendizaje automático se actualiza de forma secuencial, a diferencia del aprendizaje por lotes convencional, donde se utiliza todo el conjunto de datos de entrenamiento a la vez ("Passive Aggressive Algorithm—For big data models", 2021). Para el algoritmo aplicado a regresión tenemos matemáticamente:

$$\hat{y}_t = f(x_t) \tag{19}$$

Después de extender la predicción $\hat{y}_t$, se recibe el verdadero resultado y_t, por lo cual se sufre una pérdida instantánea basada en la discrepancia entre y_t y $f(x_t)$, el objetivo así del denominado algoritmo de aprendizaje en línea es minimizar la pérdida acumulada. Las pérdidas dependen de un parámetro predefinido ϵ y se denotan $l_\epsilon(w;(x,y))$ (Crammer et al., 2006). Para la regresión la intensidad ϵ es:

$$l_\epsilon\big(w;(x,y)\big) = f(x) = f(x) = \begin{cases} 0, & |y - w \cdot x| \leq \epsilon \\ |y - w \cdot x| - \epsilon, & otherwise \end{cases} \tag{20}$$

- ***Theil-Sen Regressio***

Esta técnica de regresión parte del estimador Theil-Sen y utiliza una generalización de la mediana en múltiples dimensiones. Por lo tanto, es robusto a valores atípicos multivariados. Aunque se debe tener en cuenta, que la robustez del estimador disminuye rápidamente con la dimensionalidad del problema. (Pedregosa et al., 2011).

Matemáticamente, la estimación del modelo se realiza calculando las pendientes y las intersecciones de una subpoblación de todas las combinaciones posibles de los p puntos de submuestra. Si se ajusta una intersección, p debe ser mayor o igual que n-

características + 1. La pendiente final y la intersección se definen entonces como la mediana espacial de estas pendientes e intersecciones (Pedregosa et al., 2011).

- **Regresión Lineal**

La regresión lineal es una de las técnicas más simples y utilizadas de algoritmos de regresión Machine Learning. Es popularmente conocida como técnica de mínimos cuadrados ordinarios, en esta se ajusta un modelo lineal con coeficientes m = (m_1, …, m_n) para minimizar la suma residual de cuadrados entre los objetivos observados en el conjunto de datos y los objetivos predichos por la aproximación lineal (Pedregosa et al., 2011). El resultado de este procedimiento se describe matemáticamente de la siguiente manera para el modelo simple:

$$y = \beta_0 + \beta_i X + \varepsilon \tag{21}$$

Por otro lado, para el modelo múltiple:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{22}$$

- ***K-Nearest Neighbors Regression***

La regresión K-Vecinos más Cercanos (KNN, por sus siglas en inglés) se puede utilizar en los casos en que las etiquetas de datos son variables continuas en lugar de discretas. La etiqueta asignada a un punto de consulta se calcula en función de la media de las etiquetas de sus vecinos más cercanos (Pedregosa et al., 2011).

Es un método no paramétrico que, de manera intuitiva, aproxima la asociación entre variables independientes y el resultado continúo promediando las observaciones en el mismo vecindario. El tamaño de la vecindad puede ser elegido a disposición o puede encontrarse mediante validación cruzada para seleccionar el tamaño que minimiza el error cuadrático medio. De esta manera se define matemáticamente, de manera más flexible como estimaciones de los vecinos, como:

$$Y = f(X) + \varepsilon$$
$$\hat{f}(x) = \frac{1}{K}\sum_{x\ in\ N} y_j \tag{23}$$

- ***Random Forest Regression***

Un bosque aleatorio es un meta estimador que se ajusta a una serie de árboles de decisión de clasificación en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el ajuste excesivo (Pedregosa et al., 2011).

El árbol de decisión es un algoritmo de aprendizaje automático supervisado utilizado tanto para problemas de clasificación como de regresión. Esta técnica crea múltiples árboles de decisión con técnicas de selección de atributos para las divisiones (o ramas), de manera muy particular. Entrenándose sobre un conjunto de datos, para posteriormente encontrar la impureza de cada nodo que es medida usando error cuadrático medio (MSE). Esto significa que el árbol trata de encontrar la división que produce las hojas tales que, en cada hoja, la variable objetivo es en promedio lo más cercano posible de las etiquetas en esa hoja en particular. Matemáticamente esto es:

$$\hat{y}_{nodo} = \frac{1}{N_{nodo}} \sum_{i \epsilon nodo} y^{(i)} \tag{24}$$

$$I(nodo) = MSE(nodo) = \frac{1}{N_{nodo}} \sum_{i \epsilon nodo} (y^{(i)} - \hat{y}_{nodo})^2 \tag{25}$$

- **Redes Neuronales Totalmente Conectadas (*Fully Connected Neural Network* (FNN))**

De manera general, denotamos matemáticamente una neurona a un nodo en donde se realizan cálculos, tomando una combinación lineal de sus entradas (axones de entrada), luego aplicar a esta combinación lineal alguna función (posiblemente no lineal) y devolver el resultado de la aplicación de la función (axón de salida). A la función que aplica la neurona se le llama función de activación. Las unidades de procesamiento se organizan en capas. Hay tres partes normalmente en una red neuronal: una capa de entrada, con unidades que representan los campos de entrada; una o varias capas ocultas; y una capa de salida, con una unidad o unidades que representa el campo o los campos de destino.

De manera particular, las Redes Neuronales Totalmente Conectadas (FNN), son una arquitectura especial de redes neuronales en donde todos sus nodos se encuentran conectados. Así pues, esto brinda unas características específicas entre las cuales encontramos:

- La principal ventaja de las redes totalmente conectadas es que son "independientes de la estructura", es decir, no es necesario hacer suposiciones especiales sobre la entrada (Ramsundar & Zadeh, 2018).
- Si bien ser independiente de la estructura hace que las redes totalmente conectadas tengan una aplicación muy amplia, dichas redes tienden a tener un rendimiento más débil que las redes de propósito especial ajustadas a la estructura de un espacio problemático (Ramsundar & Zadeh, 2018).

De manera general, la forma matemática de una salida de una capa para una red totalmente conectada es:

$$y_i = f(w_1 x_1 + \ldots + w_n x_n) \tag{26}$$

De donde la salida se visualiza como un vector formado por cada una de estas salidas, así:

$$y = \begin{pmatrix} \sigma(w_{1,1} x_1 + \cdots + w_{1,m} x_m) \\ \vdots \\ \sigma(w_{n,1} x_1 + \cdots + w_{n,m} x_m) \end{pmatrix} \tag{27}$$

Así pues, esto lo podemos visualizar como el producto punto de dos vectores, para una capa de neuronas es común denotar a y como producto matricial, como:

$$y = \sigma(wx) \tag{28}$$

5.3.4 Métricas De Evaluación

Para evaluar el desempeño de los modelos se utilizaron 3 métricas basadas en el cálculo del error o diferencia entre las variables reales y las variables predichas por los modelos, las cuales estudian la convergencia al mejor resultado de los modelos de regresión. El primer método utilizado es el MSE, definido en la ecuación (27), dónde N es el número

de muestras y_i es el resultado real y $\widehat{y_i}$ es la predicción realizada por el modelo, dentro de la ecuación se calculan las distancias y se elevan al cuadrado, lo que genera que los errores (o distancias) más altos tengan más peso dentro de esta métrica que los más bajos, dada la naturaleza de la función potencia, una desventaja del uso de esta métrica es el hecho de que devuelve unidades al cuadrado, lo cual en muchas ocasiones conduce a la imposibilidad de interpretaciones. Asimismo, el segundo método es la raíz cuadrada del error cuadrático medio (RMSE, por sus siglas en inglés), evidenciado en la ecuación (28), que, al devolver la raíz cuadrada del método anterior, nos facilita posibles interpretaciones de este error dentro del estudio del método de regresión utilizado, manteniendo la propiedad de los mayores pesos. Por último, el tercer método utilizado es el error medio absoluto (MAE, por sus siglas en inglés) representado por la ecuación (29), esta evalúa la distancia absoluta entre las observaciones (del conjunto de datos) a las predicciones en una regresión, para posteriormente calcular el promedio aritmético de estas, el uso del valor absoluto dentro de dicha ecuación permite que los errores negativos se contabilicen apropiadamente. La mayor diferencia presentada por el tercer método con respecto a los anteriores radica en que la resta entre los valores reales y predichos, estos no se elevan al cuadrado, por el contrario, se aplica el valor absoluto, de este modo, todos los errores tendrán el mismo peso en una escala lineal y no cuadrática como es el caso de los métodos anteriores.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \quad (29)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \quad (30)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \quad (31)$$

Adicionalmente para la evaluación final del modelo que mejor desempeño tenga, se adiciona la métrica coeficiente de determinación R^2, presentada en la ecuación (30) dónde MSE(pred) el cálculo presentado en la ecuación (27), y MSE(med), es el cálculo del MSE con respecto a la media presentado en la ecuación (31), dónde $\underline{y}$ representa la media de los datos. Entre más cercano esté R^2 a 1 se considera que los datos predichos se ajustan mejor a los datos reales.

$$R^2 = 1 - \frac{MSE(pred)}{MSE(med)} \quad (32)$$

$$MSE(med) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \underline{y})^2 \quad (33)$$

5.4 NIVEL DE MADUREZ TECNOLÓGICA O TRL

Uno de los logros más importantes en la industria de la ciencia de datos, es poder llevar los modelos desarrollados a producción, es decir a un desarrollo tecnológico, en este caso a un prototipo listo para pasar las fases de prueba, validación y comercialización.

Según la NASA (NASA, 2012), los niveles de madurez tecnológica (TRL) son un tipo de sistema de medición que se utiliza para evaluar el nivel de madurez de una tecnología en particular. Cada proyecto de tecnología se evalúa frente a los parámetros para cada nivel de tecnología y luego se le asigna una calificación TRL basada en el progreso del proyecto como se evidencia en la Figura 5. Hay nueve niveles de preparación tecnológica. TRL 1 es el más bajo y TRL 9 es el más alto.

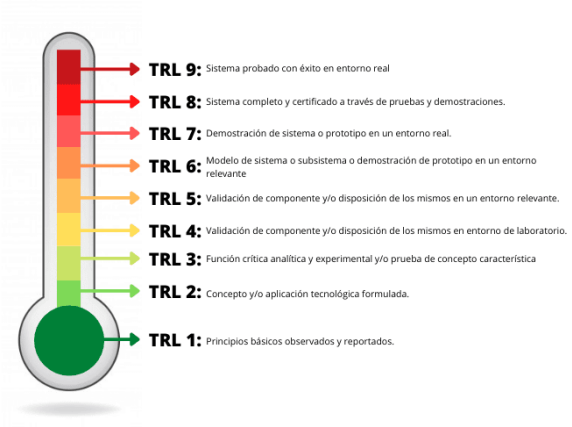
Cuando una tecnología está en TRL 1, la investigación científica está comenzando y esos resultados se están traduciendo en investigación y desarrollo futuros. TRL 2 ocurre una vez que se han estudiado los principios básicos y se pueden aplicar aplicaciones prácticas a esos hallazgos iniciales. La tecnología TRL 2 es muy especulativa, ya que hay poca o ninguna prueba de concepto experimental para la tecnología.

Cuando comienza la investigación y el diseño activos, una tecnología se eleva a TRL 3. Por lo general, se requieren estudios analíticos y de laboratorio en este nivel para ver si una tecnología es viable y está lista para continuar con el proceso de desarrollo. A menudo, durante TRL 3, se construye un modelo de prueba de concepto. Una vez que la tecnología de prueba de concepto está lista, la tecnología avanza a TRL 4. Durante TRL 4, se prueban varios componentes entre sí. TRL 5 es una continuación de TRL 4, sin embargo, una tecnología que está en 5 se identifica como una tecnología de tablero y debe someterse a pruebas más rigurosas que la tecnología que solo está en TRL 4. Las simulaciones deben ejecutarse en entornos que sean lo más realistas posible. como sea

posible. Una vez que se completa la prueba de TRL 5, una tecnología puede avanzar a TRL 6. Una tecnología TRL 6 tiene un prototipo completamente funcional o modelo de representación ya se encuentra en Ingeniería/escala piloto, validación de sistema similar en un entorno relevante: Representa un importante avance a la hora de demostrar la madurez de una tecnología. Por ejemplo, probar un sistema prototipo a escala de ingeniería con una gama de simuladores. Aquí comienza el desarrollo de ingeniería de la tecnología como sistema operativo. TRL 7 Sistema similar a gran escala demostrado en un entorno relevante el cual requiere la demostración de un prototipo de sistema real en un entorno relevante. Por ejemplo, la prueba de prototipos a gran escala con una variedad de simuladores en la puesta en marcha en frío.

TRL 8 Sistema real completado y calificado a través de prueba y demostración: se demuestra que la tecnología ha funcionado en su forma final y en las condiciones esperadas. Este TRL representa el final del verdadero desarrollo del sistema incorporando un diseño comercial. Finalmente, TRL 9 Sistema listo para su uso a escala completa: ha llegado a su forma final funcionando bajo una amplia gama de condiciones de operación. Por ejemplo, el uso del sistema real con la gama (Evalue, 2020).

Figura 5 Niveles De Madurez Tecnológica



6 OBJETIVOS

6.1 OBJETIVO GENERAL

Determinar acciones de mejora en la capacidad del proceso en la recolección de residuos sólidos urbanos integrando *Lean Six Sigma* y ciencia de datos.

6.2 OBJETIVOS ESPECÍFICOS

1. Determinar las especificaciones de calidad de la operación de recolección de residuos sólidos ordinarios bajo los criterios de LSS.
2. Caracterizar las variables internas del modelo de recolección de residuos y su impacto sobre las especificaciones de calidad del proceso a partir de la metodología DMAIC y ciencia de datos.
3. Identificar las acciones de mejora del proceso de recolección de residuos sólidos, a partir de técnicas estadísticas y ciencia de datos.

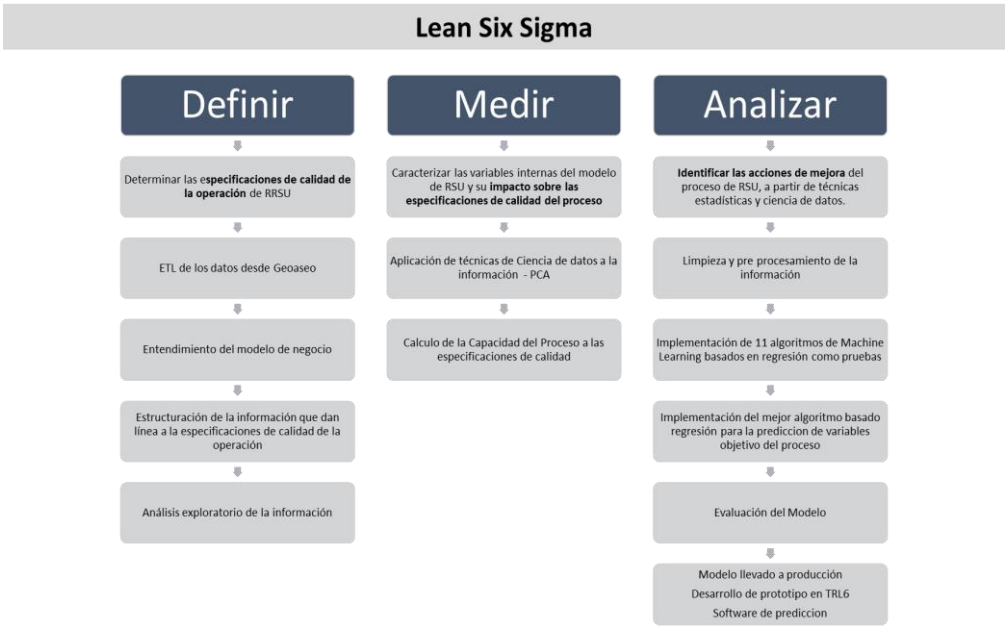
7 METODOLOGÍA

Esta investigación parte de un enfoque cuantitativo no experimental, partiendo de la identificación y caracterización de las variables del modelo de datos de una industria de aseo hasta correlacionarse por medio de técnicas de ciencia de datos.

Esta investigación es aplicada a la industria de recolección de las ciudades, utiliza metodología *Lean Six Sigma* innovando su análisis al incluir minería de datos como elemento estratégico en la selección de la relación entre las variables, pero además integra el resultado de la capacidad del proceso, para establecer la relevancia de las variables.

En la Figura 6. Se presenta el diseño metodológico de este proyecto teniendo en cuenta cada uno de los objetivos propuestos.

Figura 6 Diseño Metodológico De La Investigación



Elaboración Propia

7.1 METODOLOGÍA PARA DETERMINAR LAS ESPECIFICACIONES DE CALIDAD DE LA OPERACIÓN DE RECOLECCIÓN DE RESIDUOS SÓLIDOS ORDINARIOS BAJO LOS CRITERIOS DE LSS.

Para determinar las especificaciones de calidad de la operación de recolección de residuos sólidos ordinarios se partió de la fase DEFINIR de la metodología de LSS y DMAIC la cual se define a continuación:

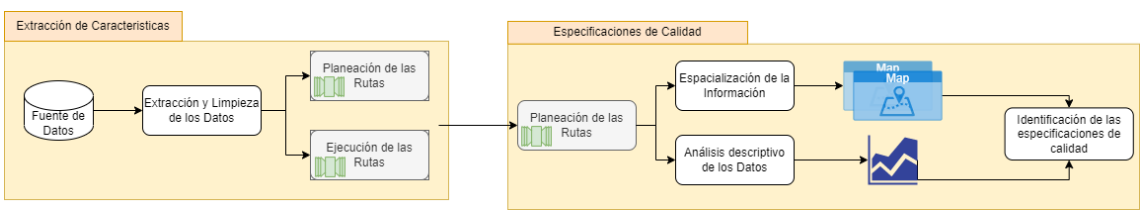
> Definir: *"Es un proceso genérico donde se define el defecto o defectos a corregir, la localización de estos, los clientes afectados, el equipo enfocado en el problema, así como los objetivos, metas y tiempos de implementación. Unidades." (Tafernaberri Franzão et al., 2016)*

Para la fase se tuvieron en cuenta los procesos de:

1. Extracción, transformación y carga de datos desde las bases de datos de GEOASEO.
2. Entendimiento del modelo de negocio en el proceso de recolección y transporte de residuos ordinarios.
3. Estructuración de la información que dan línea a las especificaciones de calidad de la operación recolección y transporte de residuos ordinarios desde las hojas de ruta planeadas y ejecutadas.
4. Análisis exploratorio de la información.

En la siguiente figura se describe el proceso llevado a cabo para dar cumplimiento al objetivo

Figura 7. Desarrollo Solución Objetivo 1



Elaboración Propia

Para este proceso se utilizaron diferentes herramientas tecnológicas desde la gestión y administración de los datos que se realiza a través de una base de datos postgres, con su componente espacial postgis, y que gestionamos la extracción de los datos a través de un administrador de consultas como pg admin (PostgreSQL, n.d.), desde esta herramienta se gestionaron todas las consultas que permitían el filtro de la información acorde a las necesidades de la gestión de la información de rutas planeadas y rutas ejecutadas.

Para un mejor entendimiento de la información espacial se visualiza la información a través de la herramienta Qgis (QGIS, n.d.), qué permite capturar información tabular con componentes geográficos y visualizar la espacialmente en los diferentes elementos geográficos de puntos líneas o polígonos.

Una vez visualizada la información y tabulada correctamente el análisis estadístico de los datos y el análisis descriptivo de los mismos se realiza a través de herramientas como Python (Python, n.d.) y Dtale, que permiten una visión cuantitativa de los datos.

7.2 METODOLOGÍA PARA CARACTERIZAR LAS VARIABLES INTERNAS DEL MODELO DE RECOLECCIÓN DE RESIDUOS Y SU IMPACTO SOBRE LAS ESPECIFICACIONES DE CALIDAD DEL PROCESO A PARTIR DE LA METODOLOGÍA DMAIC Y CIENCIA DE DATOS.

Para para caracterizar las variables internas del modelo de recolección de residuos y su impacto sobre las especificaciones de calidad del proceso se partió de la fase MEDIR de la metodología de LSS y DMAIC la cual se define a continuación:

> *Medir: La etapa de medir responde a la pregunta, ¿cuál es la capacidad de nuestro proceso?, y se puede definir como "Consiste en medir los fallos generados en aquellos procesos internos problemáticos identificados, los cuales ocasionan características críticas para la calidad del producto o servicio, es decir fuera del margen de tolerancia." (Tafernaberri Franzão et al., 2016)*

Para la fase se tuvieron en cuenta las siguientes actividades:

1. A partir de la aplicación de técnicas de ciencia de datos, como PCA se decidieron cuáles eran variables involucradas que en la base de datos Geoaseo es decir, cuales tienen más peso dentro de la correlación total.
2. Se calculó la capacidad del proceso a partir de las condiciones presentadas en el marco teórico Ítem 5.3.2, este se programó en el lenguaje de programación libre Python, donde se pudo englobar las variables del proceso que daban especificación de calidad en este caso: Toneladas Recogidas, Tiempo Total, y Kilómetros Totales.

   Para el cálculo también se tuvo en cuenta que inicialmente se tienen un total de 172 rutas ejecutadas durante diferentes días de la semana, sin embargo, para tener fiabilidad en los cálculos realizados, solo se hace el cálculo del Cp respectivo cuando la ruta, en cada uno de los días, contenga más de 100 muestras. De este modo se obtienen 103 rutas que cumplen dicha condición. Generando un procesamiento de aproximadamente 59000 registros.

7.3 METODOLOGÍA PARA IDENTIFICAR LAS ACCIONES DE MEJORA DEL PROCESO DE RECOLECCIÓN DE RESIDUOS SÓLIDOS, A PARTIR DE TÉCNICAS ESTADÍSTICAS Y CIENCIA DE DATOS.

Para identificar las acciones de mejora del proceso de recolección de residuos sólidos se partió de la fase ANALIZAR de la metodología de LSS y DMAIC, haciendo uso de técnicas estadísticas y de ciencia de datos para dar cumplimiento al objetivo.

> Analizar: Responde a la pregunta ¿Cuándo y dónde ocurren los defectos? *"Se pretende comprender el motivo por el que se producen defectos. Es usual el uso de técnicas como tormentas de ideas y herramientas estadísticas donde se identifican las variables clave. Al mismo tiempo se examinan los resultados óptimos con el fin de analizar los procedimientos que se llevaron a cabo y poder estandarizarlos." (Tafernaberri Franzão et al., 2016).*

En este objetivo se analizó, diseñó y se implementó un modelo que predice las variables en el proceso de RSS, haciendo uso técnicas de ciencia de datos, esto con

el fin de visualizar la capacidad del proceso y el comportamiento real de la operación en un nuevo módulo como sofisticación de software del producto Geoaseo.

Este se dividió en 3 fases, descritas a continuación:

Fase 1: Base de datos y preprocesamiento de la información

Para la ejecución del presente trabajo, se cuenta con una base de datos facilitada por GEOASEO. Los datos obtenidos corresponden a 80.513 registros reales de recolección de residuos sólidos en el límite urbano. Cada uno de estos registros contiene información como códigos de rutas y vehículos, fecha de recolección, día de la semana, hora de inicio y finalización, tiempo total del proceso, kilómetros totales del proceso, número de viajes requerido, número de compactaciones de los residuos sólidos, y la cantidad de combustible utilizado en toneladas.

De este modo, con la base de datos mencionada, se procede a realizar una limpieza de datos y seleccionar la información que se espera predecir y las características que se usarán para cumplir esta tarea. Basado en la presencia de valores atípicos en la información, relacionados al error humano, donde se decide eliminar estos registros, los cuales contienen valores como tiempos y toneladas negativos, dado que no tendrían sentido alguno dentro del proceso de recolección. Asimismo, la selección de información o características a utilizar se realiza haciendo un estudio cuantitativo del aporte que podría tener realizar una predicción de las características del proceso, así como, la cantidad de datos faltantes en las muestras. Estos son arrojados desde el análisis PCA resuelto en el objetivo 2.

Finalmente, se obtiene un total de 59.671 muestras, y las características seleccionadas para realizar la respectiva predicción fueron: “Tiempo Total”, “Toneladas Recogidas”, “Km Total” y “Número Compactaciones”, todas estas características representadas por valores numéricos continuos. Del mismo modo, buscando aportar características que aporten al modelo, se utilizan las variables categóricas “día”" y “número de viajes”

Fase 2: Diseño y desarrollo de modelos basados en ciencia de datos.

Dada la naturaleza de los datos se espera predecir el valor continuo de una variable basado en un conjunto de características, se plantea el uso de técnicas de aprendizaje de máquina basadas en regresión, siendo usadas las siguientes:

- *Support Vector Regression (SVR).*
- *Stochastic Gradient Descent Regression.*
- *Bayesian Ridge Regression.*
- *LassoLars.*
- *Automatic Relevance Determination Regression.*
- *Passive Aggressive Regression.*
- *Theil-Sen Regression.*
- *Linear Regression.*
- *K-Nearest Neighbors Regression (KNNR).*
- *Random Forest Regression (RFR).*
- *Fully connected neural network* - Redes Neuronales Totalmente Conectadas.

De este modo, cada uno de los modelos mencionados se aplican para la predicción de las variables: “Tiempo Total”, “Toneladas Recogidas”, “Km Total” y “Número Compactaciones”.

Para evaluar el desempeño de los modelos se utilizaron 3 métricas basadas en el cálculo del error o diferencia entre las variables reales y las variables predichas por los modelos, las cuales estudian la convergencia al mejor resultado de los modelos de regresión, estas se encuentran definidas en el Ítem 5.3.4 del marco teórico.

La metodología usada para esta predicción se basa en realizar combinaciones entre las características mencionadas, es decir, tres de las variables como entrada a los modelos y una variable como salida y puede ser dividida en 3 experimentos:

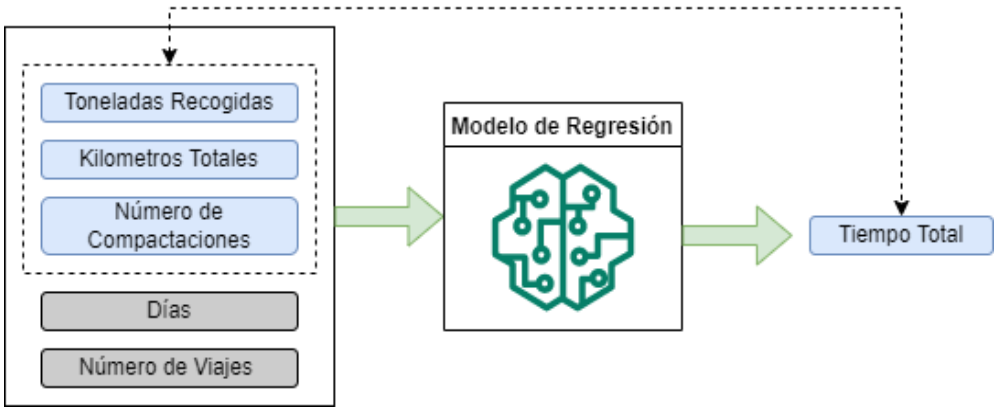
- Experimento 1: Se utilizan únicamente las variables mencionadas, con combinaciones entre ellas.
- Experimento 2: A las variables usadas en el experimento 1 se adicionan las

variables “día” y “número de viajes" como entrada a los modelos, dada la importancia que analíticamente representan en la predicción de las cuatro variables que se desea predecir.

- Experimento 3: Se mantienen las variables del experimento 2, sin embargo, se aplica una estandarización de los datos previo al ingreso de las variables al modelo, de tal forma que se disminuya el efecto que puede tener la dispersión de los datos. En este proceso, se presenta la estandarización aplicada, definida en la Ecuación (9) del marco teórico. Siendo Z el conjunto de datos estandarizado y x la muestra que se desea estandarizar.

En la Figura 8, Se puede apreciar la metodología expuesta, donde se tienen como entrada del modelo las seis variables mencionadas en los experimentos 2 y 3, y como salida la variable que se espera predecir. La distribución de los datos usada para la ejecución de la metodología es 80%-20% como conjuntos de entrenamiento y prueba, respectivamente.

Figura 8 Diseño De Modelos De Ciencia De Datos



Elaboración Propia

Fase 3: Desarrollo del módulo de ciencia de datos para sofisticación de software Geoaseo.

Por último, en esta fase del proyecto se integra el producto en un prototipo donde se articula el cálculo de la capacidad del proceso resuelto en el objetivo 2, y los modelos en ciencia de datos en un software en un nivel tecnológico en TRL6.

A esta fase del proceso se le denominó “Desarrollo del módulo de ciencia de datos para sofisticación de software Geoaseo”, el cual se llevó a producción bajo la metodológica implantación y aceptación de sistemas de información, proceso que hace parte de la Metodología Métrica V3 (*PAe - Métrica v.3*, n.d.)que se describe a continuación:

- **Análisis del sistema (ASI):** El objetivo es tener las especificaciones que respondan a las necesidades de los usuarios y puedan ser empleadas para integrar en el diseño. Disponibilidad de recursos tecnológicos como servidor para trabajar en entorno
- **Diseño del sistema (DSI):** Integración en el sistema de los requerimientos especificados en el análisis de pruebas funcionales y no funcionales.
- **Implantación (Prueba Piloto) (IAS):** Implantación es el proceso de instalación para la disposición del cliente / usuario / piloto.

Las fases de mejorar y controlar de la metodología DMAIC y LSS, no hacen parte del desarrollo del proyecto, ya que el fin es conocer el comportamiento estadístico del problema y no mejorarlo o controlarlo, de igual manera se mencionan para establecer los posibles trabajos futuros que puede desencadenar el presente proyecto de investigación

4. Mejora: *“Tiene por objetivo identificar las variables que se pueden mejorar para cuantificar el efecto sobre las características más críticas de la calidad; así en base a su relevancia, mejorar el proceso para cumplir con los márgenes aceptables.” (Tafernaberri Franzão et al., 2016)*

5. Controlar: *“En la última etapa se intenta garantizar que la modificación presente en las variables esté dentro de los márgenes de variación aceptados, se usan técnicas como el Control Estadístico de Procesos y gráficas de control.” (Tafernaberri Franzão et al., 2016)*

8 RESULTADOS

En este capítulo se presentan los resultados obtenidos para cada uno de los objetivos. Inicialmente, se presentan las especificaciones de calidad de la operación de RRS, a partir de análisis de la información que hacía parte de las hojas de rutas planeadas de la operación bajo los criterios dados por LSS. Seguidamente se presenta la caracterización de las variables internas del modelo de recolección de residuos y su impacto sobre las especificaciones de calidad del modelo del negocio desde las hojas de rutas ejecutadas, es decir la información real del proyecto. Y finalmente como acción de mejora se analizó, diseñó e implementó un prototipo en TRL6 donde se encuentran integrados 2 modelos basados en técnicas de ciencia de datos específicamente en Machine Learning para la predicción de las variables estratégicas que definen la capacidad de la operación de RRS.

8.1 ESPECIFICACIONES DE CALIDAD DE LA OPERACIÓN DE RECOLECCIÓN DE RESIDUOS SÓLIDOS ORDINARIOS BAJO LOS CRITERIOS DE LSS.

Determinar las especificaciones de calidad de las operaciones de recolección de residuos sólidos ordinarios bajo Los criterios de LSS, implica analizar las fuentes de datos de la planeación de las operaciones de aseo, el primer requerimiento importante dentro de la planeación es la obtención de la ejecución de las hojas de ruta en este instrumento se encuentra consignada la forma en que las operaciones deben ejecutar su trabajo en cada una de las rutas planeadas indicando la hora de salida, la hora de llegada los kilómetros recorridos planeados, los tiempos planeados de operación y todas las otras variables que determinan la calidad de la operación en una empresa de recolección de residuos sólidos urbanos.

Por tal motivo la primera etapa dentro del proceso implica una extracción de características de las fuentes de datos dónde se encuentra la planeación, para nuestro caso de estudio la fuente de datos se encuentra en un sistema de información geográfico que permite la planeación de las operaciones determinado como GEOASEO.

A pesar de que la información se encuentra estructurada y modelada para un sistema de información geográfico, no necesariamente cumple las especificaciones necesarias para un sistema de análisis de datos, la redundancia de la información la duplicidad de los mismos así como variables en estado vacío pueden ser elementos que afecten la calidad de la información al momento del análisis de los datos de la misma forma las planeaciones pueden tener elementos estáticos donde se estipula el inicio fin de la operación o duración de los tiempos de recolección de forma estática y no necesariamente coinciden con la realidad de la dinámica de una operación de aseo.

En la primera etapa del proceso de la extracción de características, se recurre a la fuente de datos GEOASEO, y se extrae la información de las rutas planeadas a través de una consulta directamente a las bases de datos de la misma forma se extrae la información de las rutas ejecutadas colocando unas condiciones sobre los elementos que se desean visualizar acorde a las planeaciones establecidas.

Una vez se completa la etapa de la extracción de características se pasó a la etapa de especificaciones de calidad en ella partimos de la planeación de las rutas, visualizando la información geográfica sobre las rutas planeadas e identificando el análisis descriptivo de los datos por medio de clusterización de la información estos elementos permitieron identificar lo que para el cliente significa especificaciones de calidad expresadas directamente en su planeación de rutas.

Una vez se ha identificado la fuente de datos correcta dentro del sistema GEOASEO, el primer elemento importante es la familiarización con las variables y el entendimiento de cada uno de ellos dentro del contexto del modelo de negocio segmentándolas por variables categóricas o cuantitativas como se evidencia en las Tablas 8 y 9.

Tabla 8. Definición De Variables Categóricas

Variables Categóricas		
Nombre	**Descripción**	**Contexto modelo de negocio**
mac_codigo	Código de la Macro ruta	El código de la Macro ruta hace referencia a un identificador que agrupa diferentes rutas y se Ejecutan en un día o en una frecuencia específica permite segmentar las operaciones no por características físicas similares sino por proximidad geográfica de la región al menos para la operación en estudio
mac_nombre	Nombre de la Macro ruta	El nombre de la Macro ruta se utiliza para identificarlo claramente a los equipos de trabajo internos de la compañía
hor_iniplan	Hora de inicio de la planeación de la ruta	la hora de inicio de planeación hace parte de los elementos principales del modelo de negocio indica en qué momento se debe iniciar la ruta y hace parte del cumplimiento y la especificación de calidad de la operación
hor_finpla	Hora final de planeación de la Macro ruta	de igual forma la hora de finalización de las Macro rutas hacen referencia a un elemento específico de calidad aumentar o disminuir los tiempos de horas de finalización planeadas puede indicar opciones de mejora dentro de la operación, sin contar que pueden impactar en los costos operativos de la planeación estratégica de la operación
phr_codigo	Código de la planeación,	El código de planeación específica un identificador único sobre el cual se atiende geográficamente en una zona específica de la ciudad es elemento que permite determinar los puntos de optimización de la operación para el caso de estudio de esta investigación
veh_codigo	Vehículo asociado a la Macro ruta	El código de vehículo es una variable estratégica dentro de la planeación de las operaciones permite determinar puntos de optimización del proceso y mejora, tiene asociadas información de tamaño del vehículo y capacidades.
phr_frecue	Frecuencia de la planeación.	La frecuencia de la planeación hace referencia a la intensidad con qué se hace el recorrido en una zona específica, Se Especifica indicando los días que debe ser visitada la zona para labores de recolección

Elaboración Propia

También se tienen identificadas las variables que son cuantitativas dentro del segmento de los datos como lo son.

Tabla 9. Definición Variables Numéricas

Variables Numéricas

Nombre	Descripción	Contexto modelo de negocio
phr_tonpla	Toneladas planeadas de recolección	Las coordenadas planeadas es el elemento objetivo de un sistema de información de recolección de residuos sólidos el objetivo es incrementar el número de toneladas recogidas disminuyendo Los costos asociados a su recolección
phr_kmplan	km planeados	Los kilómetros planeados es una variable que determina gran parte de Los costos asociados a la recolección en estos kilómetros planeados se encuentran relacionados la gran mayoría de los costos asociados a combustible horas de trabajo desgaste del vehículo y otro tipo de variables que impactan los costos de la operación
phr_numope	Número de operarios	El número de operarios es una de las variables que impactan los costos de operación porque se asocia a los costos de mano de obra entre menor sea la cantidad de operarios relacionados menor van a hacer los costos pero en los rendimientos en tiempos pueden ser un poco más complejos, además por las características propias de algunas zonas donde hay dificultades de acceso es necesario contar con un mayor nivel de operarios con el objetivo de cumplir los tiempos previstos de operación.
phr_galone	Galones planeados de combustible	Los galones planeados son una de las variables más fluctuantes dentro del modelo de negocio entendiendo que está determinado por variables exógenas al modelo de datos por tal motivo tener las controladas hacen referencia a un sistema que se encuentra con un nivel de madurez alto.
num_viajes	Número de viajes en los que se debe realizar la operación de recolección.	Una zona específica de la ciudad puede ser atendida en más de un viaje dependiendo del tamaño del vehículo y sus capacidades, pero además del comportamiento complejo de la zona de recolección que en ocasiones incrementa su producción de basura dependiendo los hábitos del territorio, esto significa, que cambian permanentemente viéndose afectados por eventos exógenos, (como días festivos, ferias, eventos, cambio de vocación del territorio, etc)

Elaboración Propia

Dentro de las primeras actividades de LSS, se especifica la necesidad de identificar cuál es la variable significativa para controlar en cada uno de los modelos de negocio para el caso del sistema de recolección de residuos sólidos urbanos las variables a controlar son las horas de recolección los kilómetros recorridos y las toneladas recogidas.

Se descarga dentro de la investigación el número de operarios y el tipo de vehículo, en caso del tipo de vehículo, se deja como un análisis posterior entendiendo que el tipo de vehículos puede tener connotaciones de diseño de la operación que no podemos alterar a

partir de las recomendaciones sugeridas en esta investigación. En la Tabla 10. Se presenta la forma en que se encuentra estructurada la hoja de rutas desde la planeación de la operación que permite identificar las variables estratégicas del modelo de negocio.

Tabla 10. Muestra Hoja De Rutas De Planeación

Mac_codmac	mac_nombre	hor_inipla	horfinpla	phr_codigo	veh_codigo	phr_frecue	phr_tonpla	phr_kmplan	phr_numope	phr_galone	num_viajes
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_993	EPN601	J	13	31	3	12	1
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_993	EPN601	J	13	31	3	12	2
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_994	WOV788	J	13	23	3	12	1
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_994	WOV788	J	13	23	3	12	2
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_995	WOV669	J	13	25	5	12	2
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_995	WOV669	J	13	25	5	12	1
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_996	DKV221	J	13	30	3	12	1
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_996	DKV221	J	13	30	3	12	2
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_997	EOZ066	J	13	17	4	12	1
M1	LUNES Y JUEVES DIA	6:00:00	14:00:00	PHREC_997	EOZ066	J	13	17	4	12	2

Elaboración Propia

Al final del proceso obtenemos 200 hojas de ruta planeadas que describen información asociada a operaciones urbanas, con una cantidad mínima de datos completos y que se ejecutan durante los años 2020 y 2021, se restringe también los datos planeados a hojas de recolección domiciliarias y no se tienen en cuenta la recolección de industriales dentro del análisis de la investigación tanto para las rutas ejecutadas como planeadas.

Por último y como parte de los filtros de los datos se excluyó de las hojas de rutas planeadas aquellas hojas que no cuentan con un número mínimo de 100 de datos de ejecución diligenciados de forma completa.

Tabla 11. Estadística Descriptiva General Datos Planeación

Estadística	**phr_tonplan**	**phr_kmplan**	**phr_numope**	**phr_galone**	**num_viajes**
Conteo	200	200	200	200	200
media	13.05	42.33	2.92	12.09	0.50
Desviación estándar	1.45	29.14	0.85	0.95	1.0
Valor Mínimo	5.0	17.0	1.0	6.0	1.0
q1 - 25%	13.0	25.0	2.0	12.0	1.0
q2-50%	13.0	30.0	3.0	12.0	1.0
q3-75%	13.0	44.0	4.0	12.0	2.0
Valor Máximo	25.0	200.0	5.0	20.0	2.0

Elaboración Propia

Desde la estadística descriptiva de los datos de la planeación como se muestra en la Tabla 11, se pueden evidenciar las cinco variables numéricas que se extraen de la hoja de ruta planeada como toneladas planeadas, kilómetros planeados, número de operarios, galones planeados, número de viajes. Esas variables cuentan con un conteo de registros de 200 qué superan las condiciones para el análisis de datos entre ellas rutas de origen urbano con más de 100 datos en ejecución planeadas durante el 2020 y 2021.

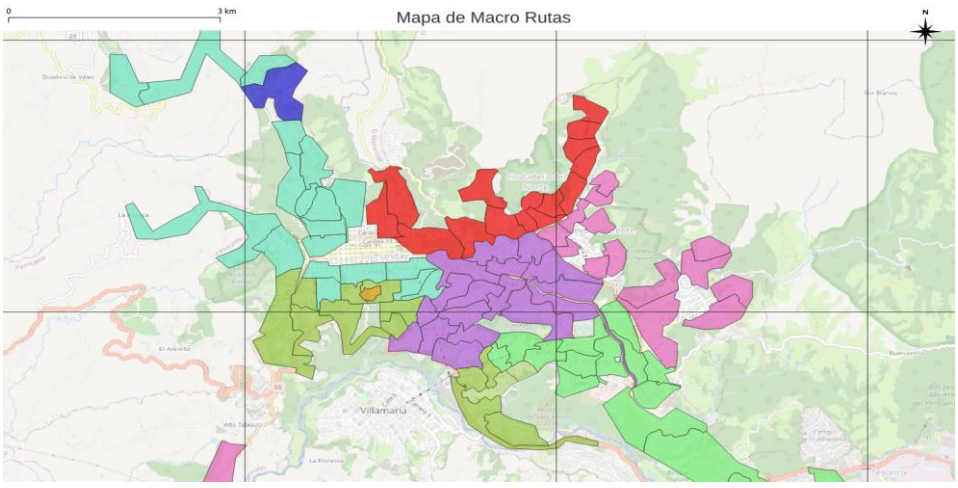
Es importante resaltar la situación que ocurre con los kilómetros planeados esta variable cuenta con límites de especificación que van desde los 17 km hasta los 200 km generando una desviación estándar del 29.14, un elemento qué evidencia dificultades para estimar los valores reales de planeación y que podrían ser asociados a la diversidad de las rutas, pero se descarta por las condiciones establecidas en la extracción de los datos.

Por otro lado, observando las toneladas planeadas se puede identificar qué la desviación estándar no es demasiado alta 1.44 pero su rango intercuartil permanece estático, lo que evidencia una planeación estática con respecto a las toneladas planeadas, condición que se repite en la variable de número de operarios, galones y número de viajes con desviaciones estándar pequeñas.

Es importante aclarar que para el modelo de negocio el número de viajes es una variable categórica porque sus valores solo oscilan entre 1 y 2 viajes.
Una vez identificados los segmentos de información requeridas y las condiciones específicas para los datos de planeación, procedemos a espacializar los datos para tener un contexto general de la cobertura de la información desde el entorno geográfico sobre el cual se está trabajando y sobre el cual contiene datos de ejecuciones de la operación.

Las Macro rutas es el primer concepto del modelo de negocio qué es necesario entender para poder abordar el análisis de la información, para esta operación la macro, agrupan zonas geográficas del territorio qué son atendidas el mismo día de operación concentrando los activos y los vehículos en un solo territorio para ganar mayor eficiencia. Como se observa en la Figura 9, se puede identificar espacios pintados de diferentes colores que muestran la segmentación de las rutas que se ejecutan en el territorio y espacio vacíos en el mapa que, para este caso, evidencian rutas que han sido ejecutadas más de 100 veces o en su defecto que han sido renombrados desde el proceso de planeación y no se encuentran las rutas ejecutadas necesarias para el análisis.

Figura 9 Diseño De Macro Rutas

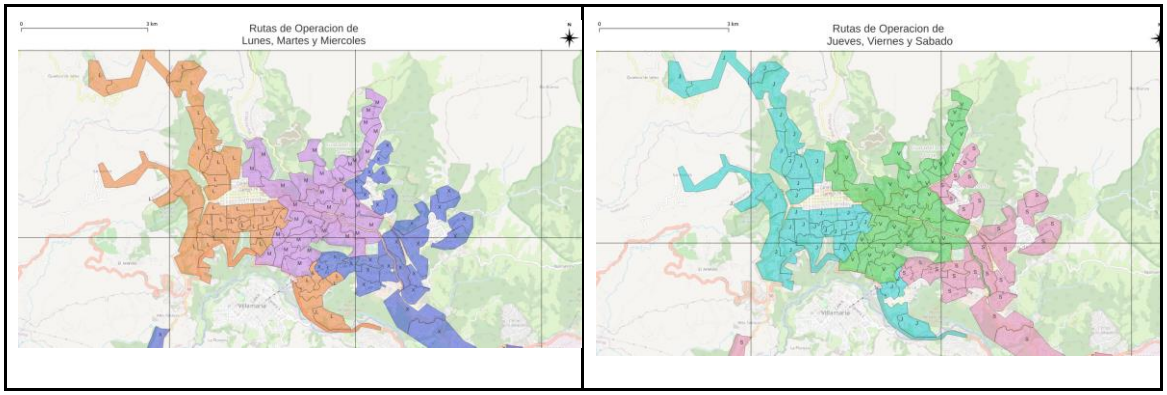


Elaboración Propia Generado por QGIS

El segundo elemento importante para entender la planeación de una operación de aseo es la frecuencia de recolección, en este caso como se observa en la Figura 10, en el mapa cobertura sobre zonas que se repiten los jueves, viernes y sábado con las operaciones que se hacen los lunes, martes y miércoles lo que indican que existe un

concepto de frecuencia sobre cada una de esas rutas expresadas en el mapa. Los polígonos al interior de cada una de las Macro rutas se conocen como rutas de recolección y son asignadas a un vehículo para ser atendidos el día indicado en la planeación de la operación

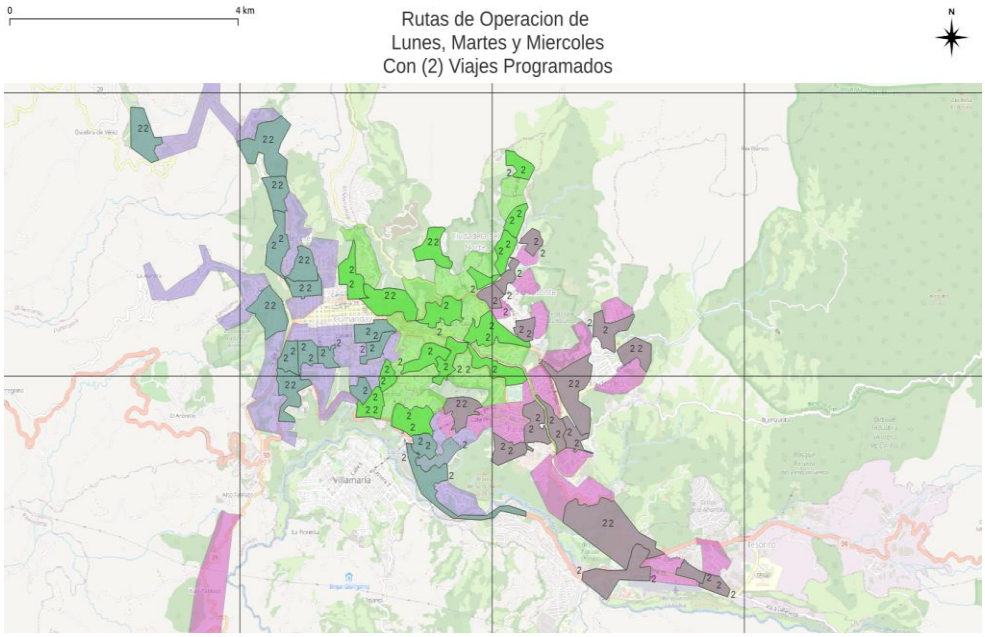
Figura 10 Concepto De Frecuencia De Recolección



Elaboración Propia Generado por QGIS

El tercer concepto importante dentro de una planeación de operación para dar total claridad a la forma y estructura de una planeación operativa es el número de viajes este concepto implica que una ruta establecida se recorre en su totalidad en una cantidad de viajes planeados que implican el desplazamiento en la zona de recolección hasta el sitio de disposición final, este procedimiento se hace tantas veces como la capacidad de la operación y la cantidad de residuos generados lo requieran, es una de las variables de la planeación de la operación

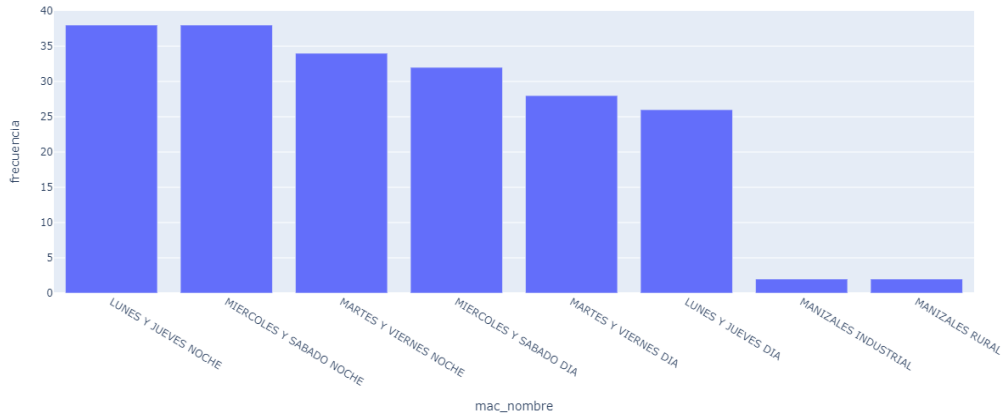
Figura 11 Concepto De Numero De Viajes



Elaboración Propia Generado por QGIS

A continuación, se presenta el análisis estadístico de la información registrada en la cobertura de los registros explicados anteriormente, es decir la base de datos denominada ‘Rutas Planeadas’, allí se consigna toda la planeación que se realiza desde la logística de la recolección de residuos sólidos en la ciudad de Manizales, en esta se cuenta con un total de 200 muestras (rutas), a estas muestras se les asocia 27 variables tales como código de macroruta (mac_codmac) y nombre de la misma (mac_nombre), código de microruta (rut_codigo), localizacion de ejecucion macroruta (mac_munici), hora inicio y fin de la ruta (hor_inipla y hor_finpla, respectivamente), sector (rut_sector), día de la ejecución (phr_frecue), toneladas planeadas (phr_tonpla), kilometros planeados (phr_kmplan), numero de galones planeados (phr_galone), número de viajes planeados (num_viajes), entre otras. Para lo cual podemos extraer algunas estadísticas descriptivas que se consideran importantes para dicho banco de datos, que nos permitan deducir algunas características, tales como:

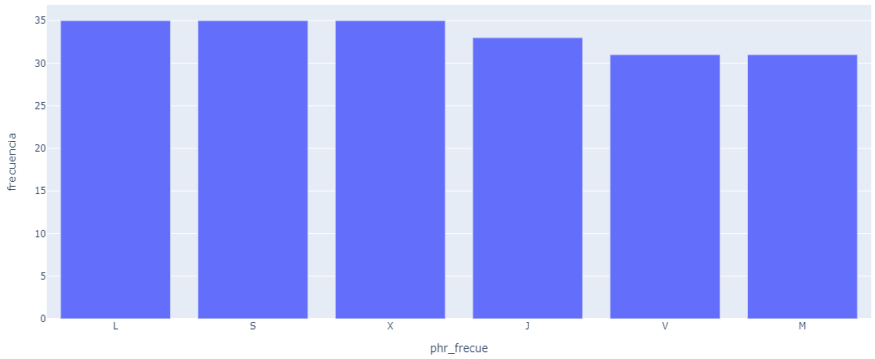
Figura 12 Cantidad De Macro Rutas Por Día De Ejecución



Elaboración Propia

En la figura 12 se evidencia La cantidad de rutas asignadas a cada una de las macro rutas existentes, por otro lado, se observa como el diseño de la operación está hecho en pares equilibrados identificando días de alta y de baja, por ejemplo, los lunes y jueves en la noche se tienen destinada la misma cantidad de rutas que se realizan el día miércoles y sábado en la noche. Este comportamiento empieza a denotar un diseño equilibrado de la operación.

Figura 13 . Cantidad De Rutas Por Días



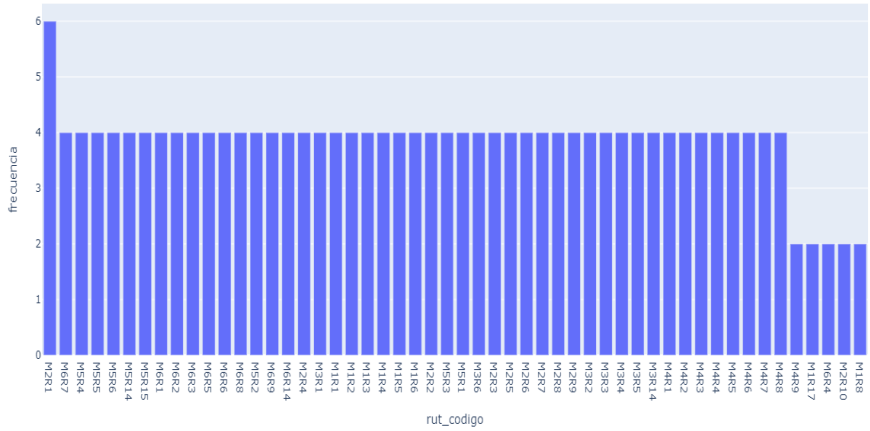
Elaboración Propia

Como se puede ver en la Figura 12, la frecuencia de las macrorutas caracterizadas por día y horario, se evidencia que para las características ‘lunes y Jueves Noche’ y

‘miércoles y Sábado Noche’ se tiene mayor cantidad de muestras que para el resto de las características de la misma variable, así mismo en la Figura 13 se observa más específicamente que los días Lunes, Sábado Miércoles y Jueves, tiene el mayor número de rutas planeadas asociadas, respectivamente.

En la Figura 14, Se puede evidenciar en primera medida la nemotecnia qué se utiliza para nombrar las rutas indicando en primera medida el número de la Macro ruta y la ruta ejecutada ejemplo M6R2, en el gráfico podemos evidenciar el diseño equilibrado de la operación a excepción de la ruta M2R1 que presenta planeaciones por encima de las cuatro frecuencias estimadas para cada una de las rutas, al final de la Gráfica podemos ver 5 rutas que tienen una frecuencia inferior y son atendidas en días específicos de la operación Pero en términos generales es una planeación equilibrada en su frecuencia de atención.

Figura 14 Cantidad De Muestras Por Código De La Ruta.

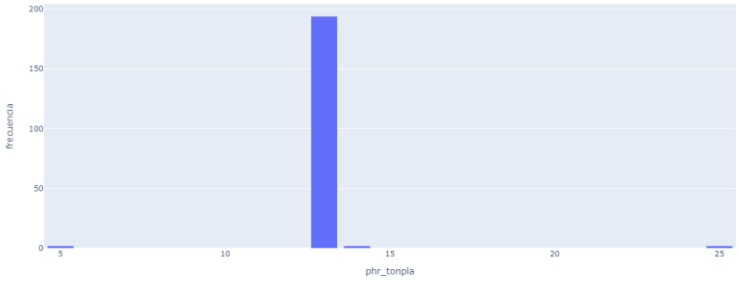


Elaboración Propia

En la Figura 15, se encuentra que a pesar con un número de más de 30 diferentes rutas, la planeación para las toneladas diverge tan solo en 4 valores numéricos, es por esto por lo que se considera a estas toneladas planeadas imprecisas e inexactas ya que no detalla de manera particular el número de toneladas. Acorde a los datos históricos de la operación.

Es importante hacer claridad que no es posible (en una etapa inicial de la operación) ajustar las toneladas recogidas más allá de una estimación por cálculo, de la cantidad de producción de residuos, asociado al número de viviendas y número de habitantes por vivienda; por tal motivo las planeaciones estratégicas tradicionales asumen una cantidad de toneladas recogidas asociadas a la capacidad del vehículo y el número de viajes que realizan en cada una de sus rutas

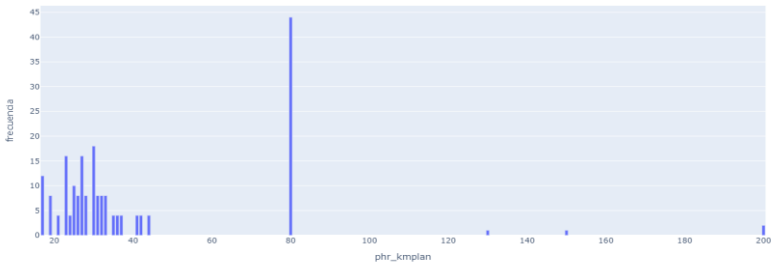
Figura 15 Cantidad De Rutas Asociadas Con Una Cantidad Específica De Toneladas Planeadas



Elaboración Propia

En la Figura 16, se evidencia un comportamiento similar con los kilómetros recorridos, a pesar de que es posible calcular los kilómetros de desplazamiento, de recolección y de transporte en cada uno de los sistemas de información geográfico que soportan las operaciones de aseo, las planeaciones están asociadas a un estándar de kilómetros por día y se ven afectadas por la percepción del diseñador de la operación, sin tener en cuenta la información histórica asociada las rutas, como proceso de actualización periódica de la planeación.

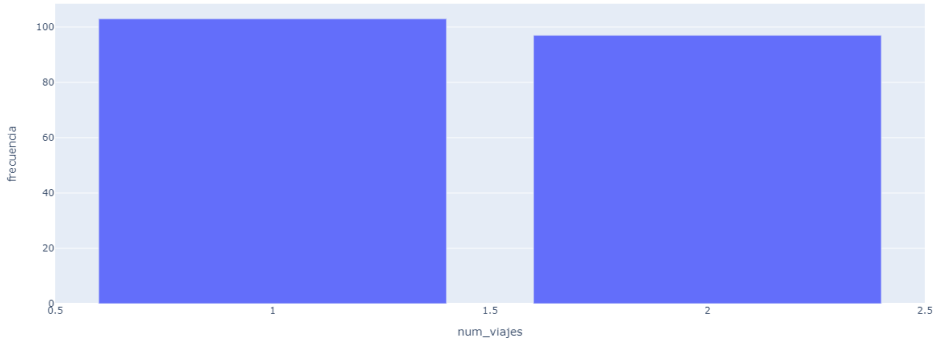
Figura 16 Cantidad De Rutas Con Una Cantidad Específica De Kilómetros Asociados



Elaboración Propia

De esta manera en la Figura 17, se evidencia que todas las rutas no cuentan con un número igual de viajes, así mismo es a partir de aquí que se deduce que los kilómetros para las rutas no se generan de manera efectiva o precisa.

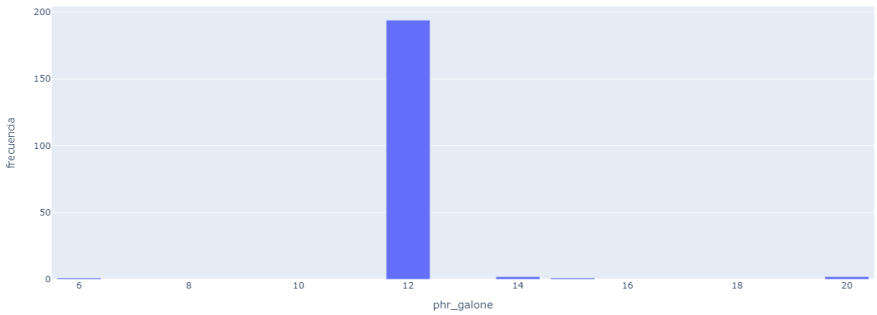
Figura 17 Cantidad De Rutas Con Un Número De Viajes Asociados



Elaboración Propia

En la Figura 18, en consecuencia, se ve el mismo comportamiento en donde a los mismos los galones requeridos planeados para la ejecución de las rutas específicas no se equipará a lo que realmente necesitaría ya que a pesar de que dichas rutas cuentan con un número de viajes y kilómetros planeados distintos, tan solo se evidencia 6 diferentes muestras que tienen un valor diferente a 12 galones para la variable phr_galone.

Figura 18 Cantidad De Rutas Asociado Con Un Número Específico De Galones Planeados.



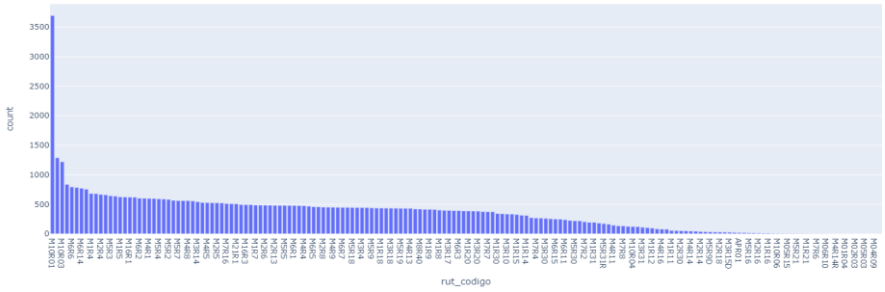
Elaboración Propia

8.2 CARACTERIZACIÓN DE LAS VARIABLES INTERNAS DEL MODELO DE RECOLECCIÓN DE RESIDUOS Y SU IMPACTO SOBRE LAS ESPECIFICACIONES DE CALIDAD DEL MODELO A PARTIR DE LA METODOLOGÍA DMAIC Y CIENCIA DE DATOS

Con el fin de caracterizar las variables internas de la operación de RSU, primero se analizaron las bases de datos de las hojas de rutas ejecutadas.

El proceso para analizar se compone de 172 rutas, como se muestra en la siguiente figura:

Figura 19 Rutas Ejecutadas



Elaboración Propia

Las cuales tienen diferente comportamiento para cada día de la semana, como se especificó en los resultados del objetivo 1.

En la sabana de datos, existen variables estratégicas para la ejecución, las cuales describen las rutas, estas son:

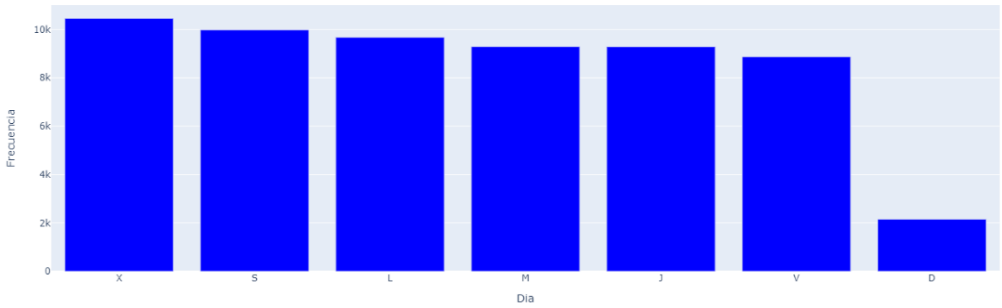
- **Tiempo Total** - Suma de todos los tiempos implicados en la ruta.
- **Toneladas recogida** - Número de toneladas recogidas en todo el recorrido hace parte de la ruta
- **Kilómetros Totales-** Kilómetros totales recorridos desde la salida del vehículo, llegada al relleno y regreso a base de operaciones.
- **Número de Compactaciones** - Número de compactaciones de los vehículos recolectores realizadas durante las recogidas.

- **Toneladas Combustibles:** Número de galones consumidos por tonelada recogida (Está variable cuenta con 6141 valores vacíos, por tanto, no se tuvo en cuenta)

A continuación, se presenta el análisis exploratorio de datos ejecutados de las variables que hacen parte del proceso:

Siguiendo el objeto de este estudio se cuenta con una segunda base de datos la cual se denomina ‘Rutas Ejecutadas’, en ella se alberga toda la información que se realiza día a día en la operación para la RRSU de la ciudad de Manizales, para esto se obtuvieron un total de 59716 registros y 35 variables, de esta cantidad de variables se consideran a rut_codigo, día, tiempo total, km total, número de viajes y tonelada recogida como las más importantes, esto debido a que son estas las que contribuirán a dislumbrar el estado actual del proceso y nos ayudará a evaluar de manera general el mismo. Asimismo, son estas variables las utilizadas para el entrenamiento de las máquinas de predicción. Por esto, en principio se extrajeron algunas estadísticas descriptivas que se consideren importantes para dicho banco de datos y nos permita deducir características.

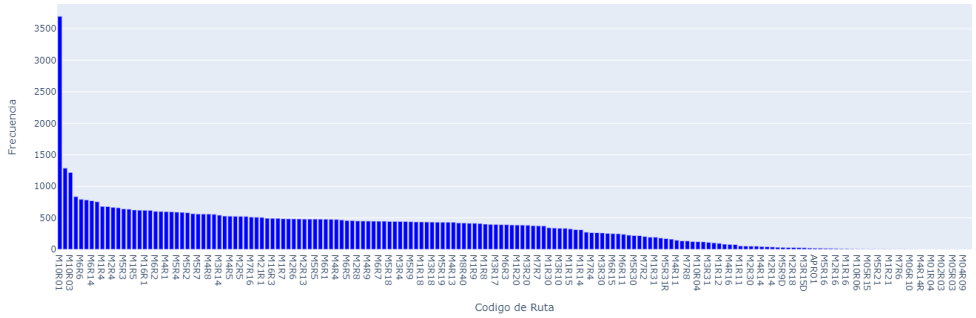
Figura 20 Cantidad De Rutas Ejecutadas Por Día De La Semana



Elaboración Propia

De la Figura 20, se puede observar que el día que se ejecutan más rutas es el miércoles, mientras que el día que menor rutas se realizan es el domingo. Es importante aclarar que todos los registros que se cuentan para esta base de datos corresponden a la cabecera municipal de Manizales.

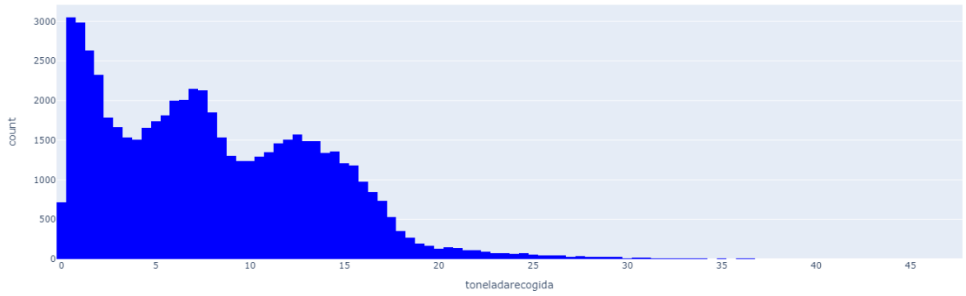
Figura 21 Frecuencia De Cada Una De Las Rutas



Elaboración Propia

De la Figura 21, se puede observar que la ruta con código M10R01 se ejecuta con mayor frecuencia contando con un total de 3695 registros asociados, mientras que rutas como M04R09, M05R03, M02R03 presentan menor número de ejecuciones dentro de la base de datos.

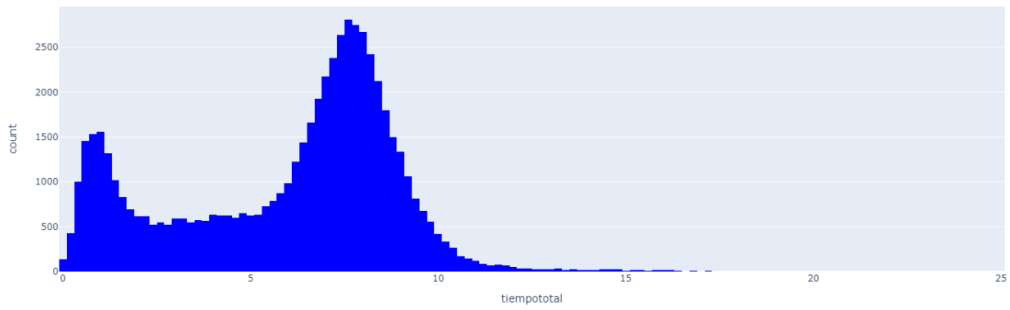
Figura 22 .Histograma Para Toneladas Recogidas Del Total De Rutas



Elaboración Propia

En la Figura 22, se visualiza la distribución de frecuencias para la variable tonelada recogida a través del histograma con 180 intervalos para el total de las rutas, como se evidencia aquí las toneladas recogidas durante la ejecución cuenta con 3 picos en frecuencia lo que nos indica un proceso no estandarizado.

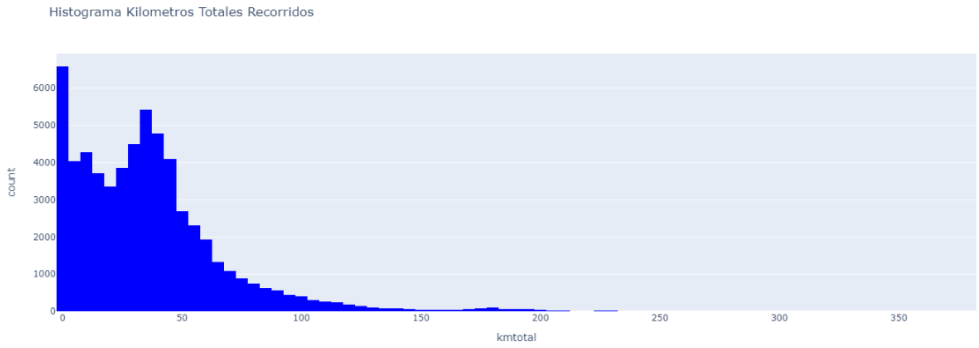
Figura 23 Histograma Del Tiempo Total Consumido Para El Total De Rutas



Elaboración Propia

Para la Figura 23, mediante la distribución de frecuencias se evidencia concentración en 2 picos principales esto nos indica que algunas rutas se ejecutan en períodos cortos de tiempo y algunas de estas en largos periodos, en contraposición a los presentado en la planeación donde se cuenta para el total de registros con tiempo total de 8 horas, así pues, se considera a la logística de este proceso como incorrecta.

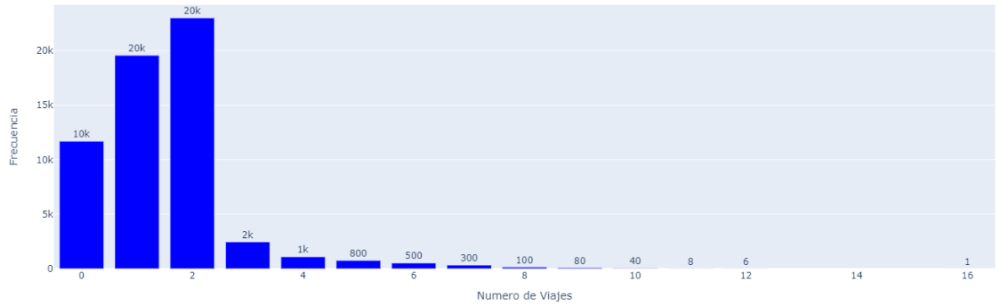
Figura 24 Histograma De Los Kilómetros Totales Recorridos Para Las Rutas.



Elaboración Propia

La distribución de frecuencias para Kilómetros Totales presentada en la Figura 24, muestra que para esta variable los registros se concentran menores a 50 kilómetros esto se puede corroborar mediante las medidas de tendencia central presentadas para esta base de datos en la Tabla 12.

Figura 25 Histograma Del Número De Viajes Para Las Rutas.



Elaboración Propia

Como se puede observar en la Figura 25, la mayoría de las ejecuciones que se realizan para la recolección conllevan más de un viaje, lo que podría conllevar a mayor consumo de combustible, desgaste del vehículo y retrasos en la finalización de la ruta.

Tabla 12. Algunas Estadísticas Descriptivas Para Las Variables Mencionadas.

	Toneladas Recogidas	**Tiempo Total**	**Kilómetros Totales**	**Número de Viajes**
Media	8.11	5.97	35.69	1.50
Desviación estándar	5.74	2.98	31.43	1.25
Valor mínimo	0	0	0	0
25%	3.17	3.63	12.50	1.00
50%	7.31	6.92	32.00	1.00
75%	12.35	8.07	47.80	2.00
Valor máximo	47.54	24.99	379.00	16.00

Elaboración Propia

Para finalizar, presentamos algunas estadísticas relevantes para las variables estudiadas anteriormente, de aquí podemos evidenciar que para las variables tonelada recogida, tiempo total, km total y número de viajes, el 75% de los datos se concentran hasta los valores 12.35, 8.07, 47.8 y 2, respectivamente. Mientras que las medias de estas variables se encuentran en 8.11ton, 5.97h, 25.69km y 1.5viajes, respectivamente.

Análisis de Componentes Principales:

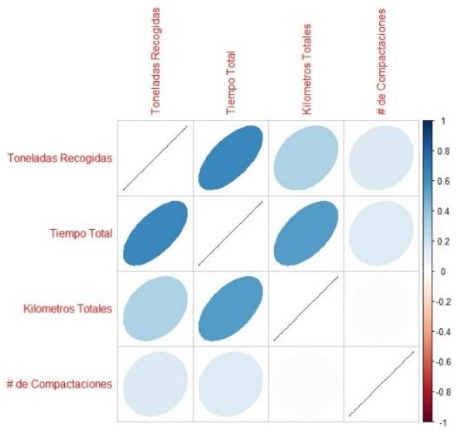
Para estudiar cuales son las variables más importantes del conjunto de datos (en otras palabras, cuales lo describen de manera más efectiva) se realizó un Análisis de Componentes Principales. En particular se tomaron las variables tiempo total, toneladas recogidas, kilómetros totales, número de compactaciones.

A partir de los resultados obtenidos en el presente análisis, se pretenden elaborar posteriores estudios con variables específicas asociadas a las tomadas en el informe. Se analizaron 59716 datos por cada una de las 4 variables implicadas para el presente análisis. Para el análisis se realizaron las siguientes 2 actividades:

1. En el primero se aplica la varianza/covarianza para estudiar la relación lineal que existe, esto se realiza cuando las variables no necesitan ser escaladas (Debido a grandes diferencias en órdenes de magnitud entre los valores de las variables a estudiar).

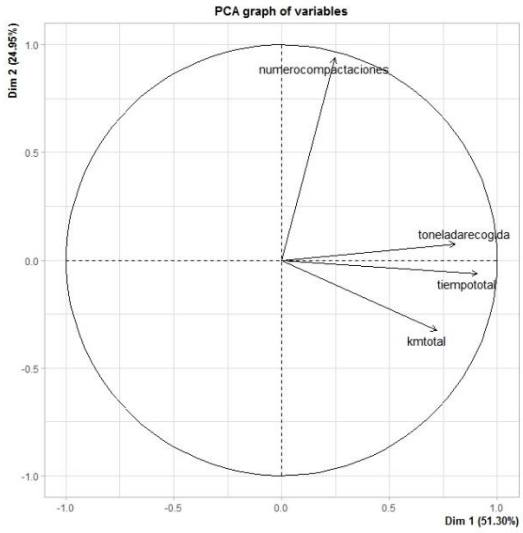
2. Para la segunda actividad se aplicó la correlación, en donde se escalan las variables para que las diferencias significativas en el rango de valores no perturben los resultados a obtener. Para las variables estudiadas aquí y en consecuencia a lo dicho anteriormente debido a que existe una brecha de órdenes de magnitud de cientos para la varianza (Figura 26), se realizará el análisis a través del estudio de la correlación, este será el camino más preciso para mejores resultados.

Figura 26 . Círculo De Correlaciones - Relación De Linealidad



Elaboración Propia

Figura 27 Peso De Las Variables En Relación Con Los Componentes Principales



Elaboración Propia.

Los resultados obtenidos del PCA (Tablas 13 y 14) se interpretan de la siguiente manera: La componente principal 1 (PC1 en la Tabla 13) está ligada a la o las variables de mayor peso (Valor(es) absoluto(s) más grande)

En este caso Toneladas Recogidas, Tiempo Total del Recorrido y Kilómetros Totales Recorridos, de donde Tiempo Total del Recorrido es la variable más importante o con mayor peso las cuales se encuentran explicando el 51.3% de todo el conjunto de datos. Por otro lado, para la segunda componente principal se tiene al Número de Compactaciones y Kilómetros Totales como la variable con mayor peso, con una explicación del 24.95% del conjunto de datos. Hasta aquí, se tiene una proporción acumulada del 76.25% para la explicación de los datos.

Tabla 13. Matriz De Varianza/Covarianza Entre Las Variables

	Tonelada recogida	Tiempo total	Kilometro total	Número compactaciones
Tonelada recogida	32.95268	11.08054	55.61479	22.34458
Tiempo total	11.08054	8.91438	52.25265	10.90021
Kilometro total	55.61479	52.25265	987.83287	10.62541
Número compactaciones	22.34458	10.90021	10.62541	670.24833

Elaboración Propia

Tabla 14. Resultados De Análisis De Componentes Principales.

	PC1	PC2	PC3	PC4
Tonelada Recogida	0.5617380	-0.07062690	0.63464290	0.52522580
Tiempo total	0.6343583	0.06319986	0.08025958	-0.76625957
Kilometro total	0.5025906	0.32640520	-0.71065019	0.36856287
Número compactaciones	0.1715884	-0.94002595	-0.29285337	0.03384587

Elaboración Propia

Tabla 15. Algunos Resultados Importantes De PCA

	PC1	PC2	PC3	PC4
Desviación Estándar	1.432	0.9991	0.8168	0.5318
Proporción de Varianza	0.513	0.2495	0.1668	0.0707
Proporción Acumulada	0.513	0.7625	0.9293	1.0000

Elaboración Propia

De lo anterior es importante entender que se tienen a Toneladas Recogidas, Tiempo Total del Recorrido y Kilómetros Totales Recorridos como las variables que explican el 51.30% de la variabilidad del conjunto de datos tomados, por ende, se recomienda el enfoque de mejora en estas tres variables para la optimización de rutas, específicamente, es la variable Tiempo Total en la que se debería enfocar los esfuerzos de mejora.

Se recomienda un estudio posterior, por ejemplo, de importancia relativa, más específico de la variable Número de Compactaciones para entender la relación que existe con las rutas y por qué esta tiene un mayor peso en la Componente Principal número 2 la cual explica el 24.95% de la variabilidad de los datos.

Se propone un método de imputación de datos para los valores perdidos en la variable Número de Compactaciones con lo cual se podría estudiar su importancia dentro de los objetivos de mejora. A partir de un estudio detallado del análisis de la capacidad de cada una de las rutas o por días se podría proponer un modelo de regresión óptimo en aras de optimizar el proceso.

Teniendo en cuenta los resultados obtenidos en el PCA, se caracterizó cuáles de las variables son más representativas, así los esfuerzos se realizan en pro a la mejora continua de dichas variables dentro de la base de datos, partiendo del cálculo de la Capacidad del Proceso, en donde diversos autores parten del supuesto de normalidad en dichas variables (Gutiérrez Pulido & de la Vara Salazar, 2009), en el presente trabajo se pretenderá observar el cálculo de la capacidad del proceso sino que también se ve involucrada en las máquinas de predicción, en este sentido se realiza otro proceso el cual es la normalización por Z-Score para el entrenamiento de los modelos, en aras de asegurar dicha normalidad.

Capacidad Del Proceso:

Para el desarrollo de este proyecto se procedió a calcular la capacidad del proceso, que se define por la fórmula explicada en el marco teórico (Ecuación 7).

En el presente caso se realiza sobre las características que demostraron tener una mayor influencia sobre el proceso: Toneladas Recogidas, Tiempo Total, y Kilómetros Totales, como se mencionó anteriormente.

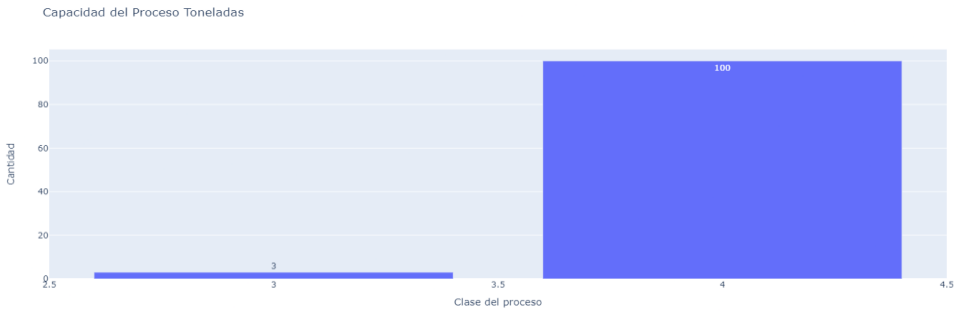
Antes de realizar este cálculo, es necesario establecer diferentes condiciones tanto generales, como para cada una de las características sobre las que se realizará el cálculo. Inicialmente se tienen un total de 172 rutas ejecutadas durante diferentes días de la semana, sin embargo, para tener fiabilidad en los cálculos realizados, solo se hace el cálculo del CP respectivo cuando la ruta, en cada uno de los días, contenga más de 100 muestras. De este modo se obtienen 103 rutas que cumplen dicha condición.

Para establecer los valores LSE y LIE, se parte de las planeaciones realizadas por los profesionales respectivos con anterioridad, siendo esta planeación diferente dependiendo de la ruta y el día en el que se realice. Ya que dichas planeaciones pueden variar, se establece que para cada ruta y día el LSE y LIE será la media de lo planeado más o menos un valor definido para que cubra un rango de valores.

CP Toneladas Recogidas: En la siguiente figura se puede apreciar el cálculo del Cp, dónde el LSE y LIE fue definido como la media de lo planeado más y menos 2

respectivamente. De las 103 rutas, basado en los valores de Cp, 100 quedaron definidas dentro de una clase de proceso 4 y las 3 restante en la clase de proceso 3.

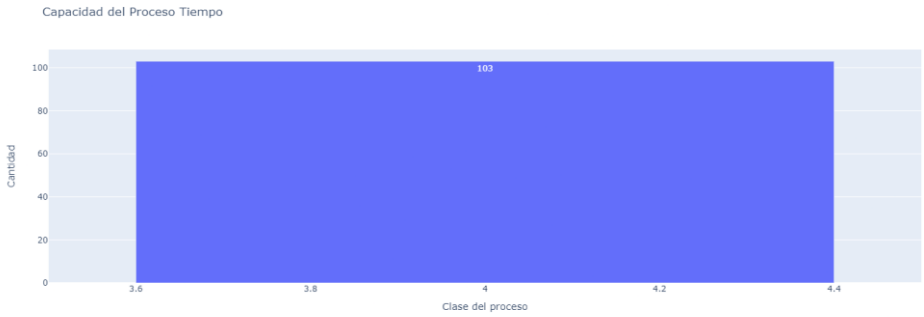
Figura 28 Capacidad Del Proceso Toneladas



Elaboración Propia

CP para Tiempo Total: En este caso, el LSE y LIE, se define por los límites de las jornadas laborales, siendo estas 8 y 6 horas respectivamente. En la siguiente figura se pueden apreciar los resultados obtenidos, las 103 rutas quedaron definidas dentro de una clase del proceso 4.

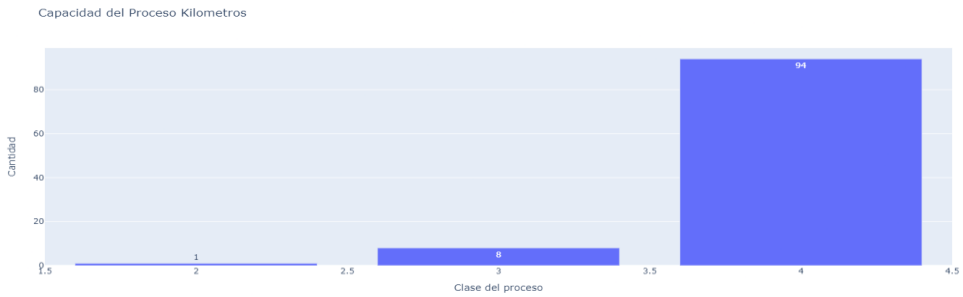
Figura 29 Capacidad Del Proceso Tiempo



Elaboración Propia

Cp para los kilómetros totales: Al igual que en toneladas los límites están definidos por la media de los datos, pero en este caso el rango es un poco más amplio, de más y menos 5, dado que se tienen en cuenta momentos en que las rutas se vean modificadas por motivos que no se pueden controlar. En este caso, las clases del proceso asignadas son 94 rutas en clase de proceso 4, 8 en clase de proceso 3, y 1 en clase de proceso 2.

Figura 30 Capacidad Del Proceso Kilómetros



Elaboración Propia

8.3 IDENTIFICAR LAS ACCIONES DE MEJORA DEL PROCESO DE RECOLECCIÓN DE RESIDUOS SÓLIDOS, A PARTIR DE TÉCNICAS ESTADÍSTICAS Y CIENCIA DE DATOS.

Cómo se expresó en la sección 7.3, en **la fase 1**, se realizó el preprocesamiento de los datos y en análisis exploratorio de los datos explícito en los resultados de los objetivos 1 y 2.

Para la **Fase 2: Diseño y desarrollo de modelos basados en ciencia de datos,** se plantearon 3 experimentos para la predicción de diferentes variables. En esta sección se mostrarán los resultados obtenidos por cada modelo, en cada una de las variables a predecir para cada uno de los experimentos. Finalmente, se mostrarán los resultados de la predicción realizada con el modelo que mejor desempeño tuvo en cada una de las variables.

- Experimento 1: en las Tablas 16, 17, 18 y 19 se pueden apreciar los resultados para la predicción de las cuatro variables deseadas, usando los diferentes modelos de regresión. De este modo, se observa cómo en todos los casos el modelo que mejor desempeño tuvo fue el basado en redes neuronales, sin embargo, es importante resaltar que los errores más grandes se encuentran en las variables “Km total" y “número de compactaciones” para las 3 métricas usadas.

Tabla 16. Resultados Tiempo Total Experimento 1

Modelo	MSE	RMSE	MAE
SVR	2.548249123	1.596323627	1.043970009
SGDRegressor	5.445184193	2.333491845	1.648536238
BayesianRidge	4.070592284	2.017570887	1.527564686
LassoLars	8.956506383	2.992742285	2.487865398
ARDRegression	4.100877765	2.02506241	1.536104065
PassiveAggressiveRegressor	10.33972694	3.215544579	2.607174148
TheilSenRegressor	5.348064804	2.312588334	1.491291628
LinearRegression	4.070597166	2.017572097	1.527553221
KNeighborsRegressor	2.558750391	1.59960945	1.061721701
RandomForestRegressor	2.508292787	1.583759069	1.037655576
Red Neuronal	**2.288676605**	**1.51283727**	**0.9828340236**

Elaboración Propia

Tabla 17. Resultados Toneladas Recogidas Experimento 1

Modelo	MSE	RMSE	MAE
SVR	17.58259303	4.193160267	2.848276492
SGDRegressor	37.22796106	6.10147204	4.046751214
BayesianRidge	19.36295048	4.400335269	3.168828832
LassoLars	33.11679802	5.75471963	4.738911899
ARDRegression	19.362755	4.400313057	3.168934794
PassiveAggressiveRegressor	154.4479302	12.42770816	9.480238198
TheilSenRegressor	20.62842817	4.541852945	3.100729948
LinearRegression	19.36303954	4.400345389	3.168766695
KNeighborsRegressor	18.22059678	4.268559099	2.888901486
RandomForestRegressor	17.90773654	4.231753365	2.844075256
Red Neuronal	**16.68251937**	**4.084423994**	**2.653138613**

Elaboración Propia

Tabla 18. Resultados Km Total Experimento 1

Modelo	MSE	RMSE	MAE
SVR	666.7739498	25.82196642	15.58630974
SGDRegressor	759.2355843	27.55422988	17.63110794
BayesianRidge	695.8309644	26.37860808	16.18530882
LassoLars	997.8334071	31.58850118	22.50983456
ARDRegression	695.8234906	26.37846642	16.18600637
PassiveAggressiveRegressor	1238.450202	35.19162119	27.09648526
TheilSenRegressor	728.4001717	26.98888978	16.35435786
LinearRegression	695.8358811	26.37870128	16.18532162
KNeighborsRegressor	674.8451333	25.97778153	15.83990534
RandomForestRegressor	658.8643073	25.66835225	15.63350633

Red Neuronal	**599.2021085**	**24.47860512**	**14.62142904**

Elaboración Propia

Tabla 19. Resultados Número de Compactaciones Experimento 1

Modelo	MSE	RMSE	MAE
SVR	563.4699706	23.73752242	17.01985541
SGDRegressor	825.7883735	28.73653378	24.46935774
BayesianRidge	655.2313496	25.59748717	22.73925854
LassoLars	676.4811211	26.00925068	23.53549512
ARDRegression	655.2329928	25.59751927	22.73932061
PassiveAggressiveRegressor	1031.122585	32.11109754	24.05205171
TheilSenRegressor	684.5643761	26.16418117	22.67227118
LinearRegression	655.233093	25.59752123	22.73804759
KNeighborsRegressor	529.1110415	23.00241382	16.31856999
RandomForestRegressor	486.542863	22.05771663	15.95897019
Red Neuronal	552.1867972	23.49865522	15.09950445

Elaboración Propia

- Experimento 2: en las Tablas 20, 21, 22 y 23, se pueden apreciar los resultados presentados para el experimento 2. De manera similar al experimento 1 el mejor desempeño lo presentan las Redes Neuronales Totalmente Conectadas (FNN), presentando mayores errores en las variables “Km total” y “Número de Compactaciones". Sin embargo, los tres errores calculados disminuyen, siendo notable el caso de la variable “Toneladas Recogidas", dónde la disminución del MAE es de aproximadamente 0,8266, teniendo en cuenta la escala de la variable.

Tabla 20. Resultados Tiempo Total Experimento 2

Modelo	MSE	RMSE	MAE
SVR	2.354437012	1.534417483	1.012395739
SGDRegressor	5.458062451	2.336249655	1.797934508
BayesianRidge	3.934182486	1.983477372	1.504818094
LassoLars	8.956506383	2.992742285	2.487865398
ARDRegression	3.933981099	1.983426605	1.504844197
PassiveAggressiveRegressor	13.18859761	3.631610884	2.474455724
TheilSenRegressor	4.996889635	2.235372371	1.443933428
LinearRegression	3.934179684	1.983476666	1.504809853
KNeighborsRegressor	2.170251887	1.47317748	0.9771471822
RandomForestRegressor	2.057545049	1.434414532	0.9336788382

Red Neuronal	**1.985370297**	**1.409031688**	**0.907381339**

Elaboración Propia

Tabla 21. Resultados Toneladas Recogidas Experimento 2

Modelo	MSE	RMSE	MAE
SVR	12.1873102	3.49103283	2.410903349
SGDRegressor	9.89860955	9.949175.6	7.305533
BayesianRidge	13.2051914	3.63389479	2.6180732
LassoLars	33.11679802	5.75471963	4.738911899
ARDRegression	13.20515223	3.633889408	2.61806361
PassiveAggressiveRegressor	31.47544124	5.610297785	4.129861655
TheilSenRegressor	14.31460711	3.783464961	2.641262147
LinearRegression	13.20524737	3.633902499	2.618059192
KNeighborsRegressor	10.82216909	3.289706536	2.191833743
RandomForestRegressor	8.663937266	2.943456687	1.881529576
Red Neuronal	**8.348805018**	**2.889429878**	**1.826447737**

Elaboración Propia

Tabla 22. Resultados Km Total Experimento 2

Modelo	MSE	RMSE	MAE
SVR	655.9237785	25.61100893	15.33758105
SGDRegressor	695.1792323	26.36625177	15.94337924
BayesianRidge	670.1037811	25.88636284	15.70555901
LassoLars	997.8334071	31.58850118	22.50983456
ARDRegression	670.1051287	25.88638887	15.70697721
PassiveAggressiveRegressor	8388.589022	91.58924076	71.28006021
TheilSenRegressor	685.0385784	26.17324165	15.73954992
LinearRegression	670.1041601	25.88637016	15.70556783
KNeighborsRegressor	615.9249179	24.81783467	14.86913955
RandomForestRegressor	600.3614882	24.50227516	14.53332145
Red Neuronal	**588.0096368**	**24.24891001**	**14.12023137**

Elaboración Propia

Tabla 23. Resultados Número de Compactaciones Experimento 2

Modelo	MSE	RMSE	MAE
SVR	537.2866755	23.17944511	16.76641169
SGDRegressor	5753.980039	75.8549935	53.96202054
BayesianRidge	597.9059017	24.45211446	21.08604743
LassoLars	676.4811211	26.00925068	23.53549512
ARDRegression	597.91291	24.45225777	21.08725503
PassiveAggressiveRegressor	9602.825015	97.99400499	75.08425308

TheilSenRegressor	627.9368265	25.05866769	21.10324995
LinearRegression	597.909681	24.45219174	21.08559057
KNeighborsRegressor	476.8647488	21.83723309	15.15210985
RandomForestRegressor	422.013271	20.54296159	14.20427579
Red Neuronal	**478.7479058**	**21.88030863**	**13.5964893**

Elaboración Propia

- Experimento 3: al igual a los experimentos 1 y 2, en las tablas 24, 25, 26 y 27. Se pueden apreciar los resultados del experimento 3, dónde el mejor desempeño es presentado por el modelo basado en Redes Neuronales Totalmente Conectadas, con mayores errores en las variables "Km total'" y "Número de Compactaciones". Sin embargo, los tres errores calculados continúan disminuyendo en comparación a los presentados en el experimento 2.

Tabla 24. Resultados Tiempo Total Experimento 3

Modelo	MSE	RMSE	MAE
SVR	2.003064762	1.415296705	0.9221247963
SGDRegressor	3.931873773	1.982895301	1.510163173
BayesianRidge	3.934157605	1.9834711	1.504843711
LassoLars	8.956506383	2.992742285	2.487865398
ARDRegression	3.933980236	1.983426388	1.504844292
PassiveAggressiveRegressor	8.719230534	2.952834322	1.890047657
TheilSenRegressor	4.666174322	2.160132941	1.449698891
LinearRegression	3.934179684	1.983476666	1.504809853
KNeighborsRegressor	2.076932133	1.441156526	0.9531602385
RandomForestRegressor	2.047678804	1.43097128	0.9323340184
Red Neuronal	**1.964200262**	**1.401499291**	**0.8934816426**

Elaboración Propia

Tabla 25. Resultados Toneladas Recogidas Experimento 3

Modelo	MSE	RMSE	MAE
SVR	9.216131025	3.035808134	1.961935937
SGDRegressor	13.21480916	3.635217897	2.616227603
BayesianRidge	13.20518183	3.633893481	2.618079725
LassoLars	33.11679802	5.75471963	4.738911899
ARDRegression	13.20515202	3.633889379	2.618063703
PassiveAggressiveRegressor	14.64202003	3.826489257	2.74234086
TheilSenRegressor	14.1676887	3.763999031	2.63265539
LinearRegression	13.20524737	3.633902499	2.618059192

KNeighborsRegressor	9.245759852	3.040684109	1.95559408
RandomForestRegressor	8.651487346	2.94134108	1.881956021
Red Neuronal	**8.483943306**	**2.912720945**	**1.814051479**

Elaboración Propia

Tabla 26. Resultados Km Total Experimento 3

Modelo	MSE	RMSE	MAE
SVR	611.7569075	24.73372005	14.71743378
SGDRegressor	671.2555028	25.90859901	15.70254366
BayesianRidge	670.0958112	25.8862089	15.70555584
LassoLars	997.8334071	31.58850118	22.50983456
ARDRegression	670.105078	25.88638789	15.70698487
PassiveAggressiveRegressor	701.8611484	26.49266216	17.12822815
TheilSenRegressor	684.4954547	26.16286404	15.5798597
LinearRegression	670.1041601	25.88637016	15.70556783
KNeighborsRegressor	616.0842795	24.82104509	14.89262101
RandomForestRegressor	597.9087532	24.45217277	14.51794228
Red Neuronal	**573.7967343**	**23.95405465**	**13.90875731**

Elaboración Propia

Tabla 27. Resultados Número de Compactaciones Experimento 3

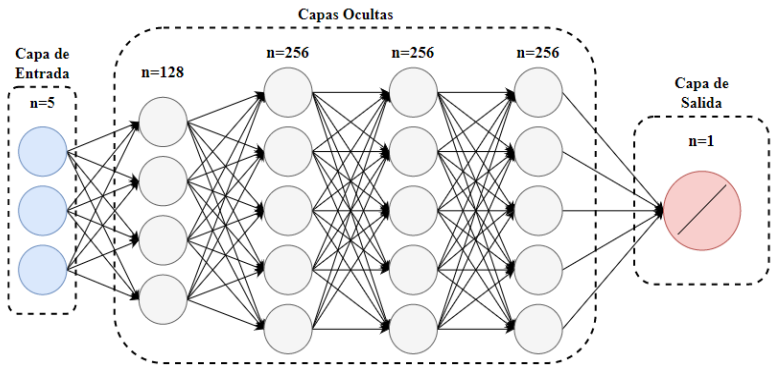
Modelo	MSE	RMSE	MAE
SVR	472.7551154	21.74293254	14.97886869
SGDRegressor	597.9575426	24.4531704	21.06346612
BayesianRidge	597.9066512	24.45212979	21.08630868
LassoLars	676.4811211	26.00925068	23.53549512
ARDRegression	597.91286	24.45225675	21.08725983
PassiveAggressiveRegressor	853.6525764	29.21733349	23.14542361
TheilSenRegressor	627.4446086	25.04884446	21.24358992
LinearRegression	597.909681	24.45219174	21.08559057
KNeighborsRegressor	457.8046919	21.396371	14.53752512
RandomForestRegressor	421.616862	20.53331103	14.20238614
Red Neuronal	**469.0609188**	**21.65781427**	**13.1763132**

Elaboración Propia

Dados los resultados presentados, el experimento que mejor desempeño tuvo fue el 3 usando el modelo de Redes Neuronales totalmente conectadas, por esta razón se procede a realizar un entrenamiento más extenso y ajustando hiper-parámetros del mismo modelo. La red neuronal final propuesta se compone de una capa de entrada,

cuatro capas ocultas, y capa de salida; las capas ocultas se componen de 128 neuronas para la primera capa y 256 neuronas para las restantes, con una función de activación unidad lineal rectificada (ReLu) en todas estas. Del mismo modo, la capa de salida se compone de una única neurona, con una función de activación lineal, debido a la tarea de regresión que se desea realizar. Finalmente, se usa cómo función de pérdida el MAE, y como algoritmo de optimización Adam. La estructura anteriormente mencionada se puede encontrar en la Figura 31, dónde N corresponde al número de neuronas presentes en cada una de las capas.

Figura 31 Estructura De La red Neuronal Emplea*da*



Elaboración Propia

De este modo, se procede a realizar el entrenamiento de la red neuronal durante 100 épocas, teniendo en cuenta además que para la predicción de las variables “Tiempo Total" y “Toneladas Recogidas” se elimina la variable de entrada “Número Compactaciones" debido a que esta variable no es posible obtenerla para hacer una predicción previo a la puesta en marcha del proceso de recolección de basuras. Cómo se evidencia en la Tabla 30, los resultados para la predicción de cada variable son muy similares a los presentados en el experimento 3, incluso teniendo en cuenta la eliminación de una variable de entrada en las predicciones de las 2 primeras variables en el cuadro. Asimismo, de tal forma que se realice una evaluación más estricta con el modelo final, se obtiene como métrica adicional el coeficiente de determinación R^2, el cual da un resultado considerablemente más alto en las 2 primeras variables, siendo este superior a 0.70; por el contrario, en las variables restantes este valor no sobrepasa en 0.50.

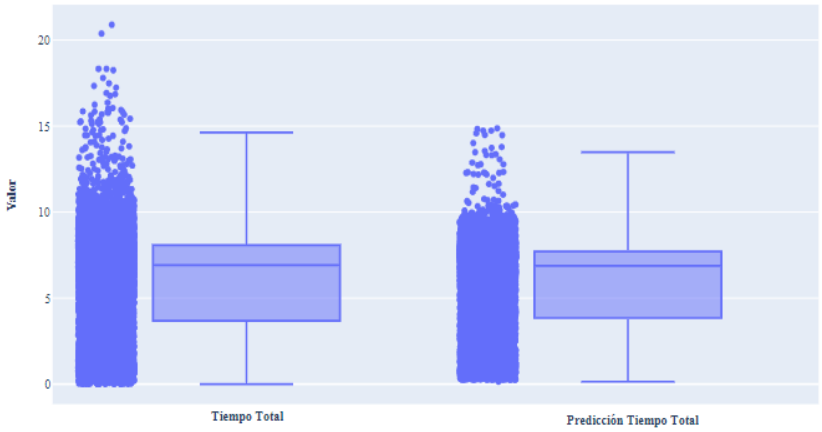
Tabla 28. Resultados Entrenamiento Con Red Neuronal

Target	MSE	RMSE	MAE	R^2
Tiempo Total	**2.0292**	**1.4245**	**0.9149**	**0.7910**
Toneladas Recogidas	8.9590	2.9931	1.9305	0.7294
Km Total	583.0044	24.1454	13.8843	0.4157
Número Compactaciones	463.5823	21.5309	13.0240	0.3146

Elaboración Propia

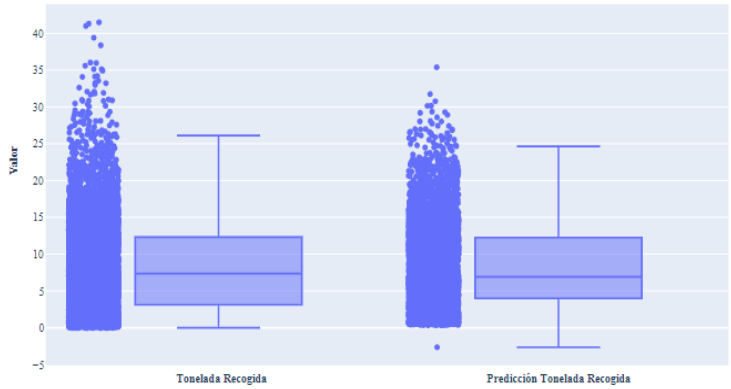
Finalmente, se realizan los diagramas de cajas de tal forma que se pueda comparar la distribución de los datos reales contra los datos que el modelo predice. Como se puede apreciar en las Figura 32 y 33 para las variables “Tiempo Total” y “Tonelada Recogida", respectivamente, las predicciones sobre el conjunto de datos tienen un comportamiento muy similar a los valores reales, siendo notable que la mediana es similar en ambos y además, las posiciones el rango entre el cuartil 1 y el cuartil 3 es similar en ambos. Por otro lado, en las Fig. 34 y 35, se evidencian los diagramas de caja para las variables “Km Total" y “Número Compactaciones", las cuales presentan una distribución con una mayor diferencia entre los valores predichos y los reales, principalmente en la Fig. 35, teniendo en cuenta, además, la escala de los valores.

Figura 32 Diagrama De Caja Tiempo Total vs Predicción Tiempo Total



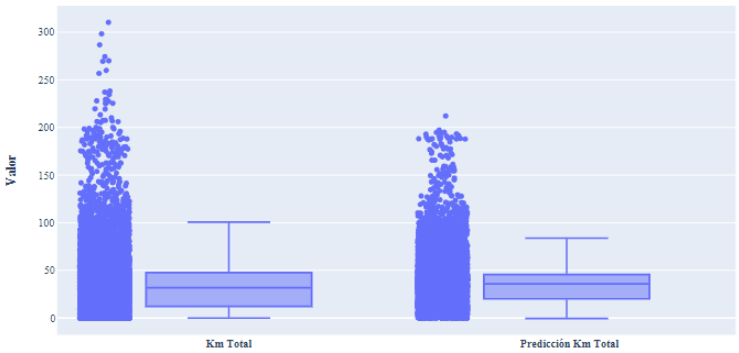
Elaboración Propia

Figura 33 Diagrama de Caja Toneladas Recogidas vs Predicción Toneladas Recogidas



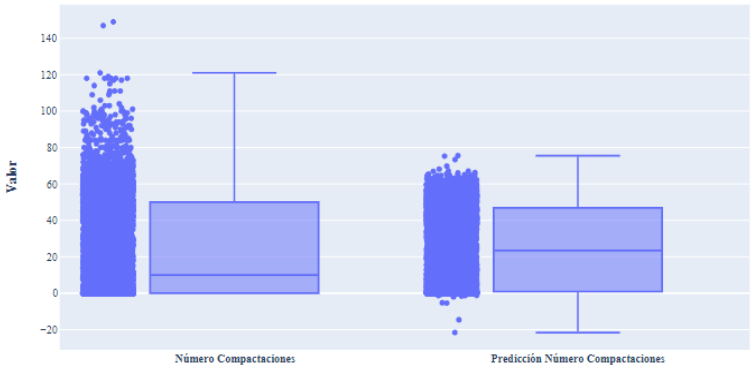
Elaboración Propia

Figura 34 Diagrama De Caja Km Total vs Predicción Km Total



Elaboración Propia

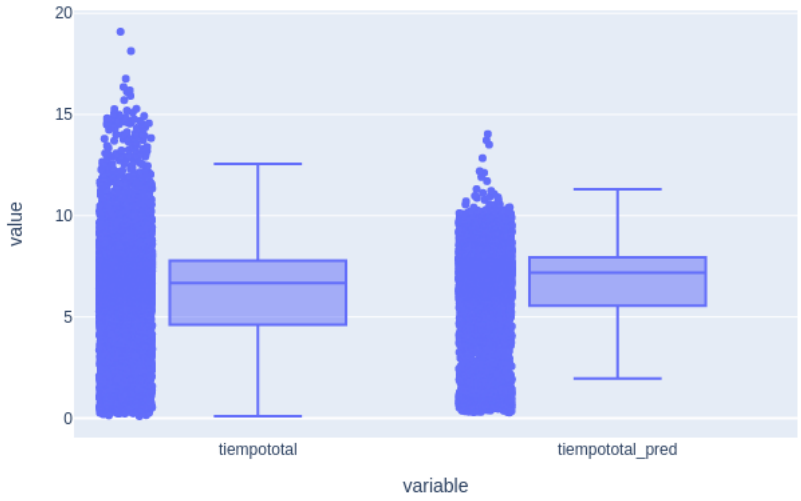
Figura 35 Diagrama De Caja Número Compactaciones vs Predicción Número Compactaciones



Elaboración Propia

Para verificar la funcionalidad final de los modelos y su consistencia en el tiempo, se realizan predicciones tanto para “Toneladas Recogidas” como “Tiempo Total”, sobre 5589 datos nuevos, recolectados entre el 01/01/2022 y el 30/05/2022. Como se puede apreciar en la Figura 36, la distribución de los datos reales, comparado con los datos de la predicción realizada por el modelo. Adicionalmente a esto, se realiza el cálculo del valor R^2 de tal forma que se estime qué tan bien se ajustan las predicciones a los datos reales. En este caso, se obtiene un valor R^2 de 0.7132, lo cual indica que los datos se siguen ajustando bien para esta característica a pesar del alto margen del tiempo en el que se realizan predicciones.

Figura 36 Resultados Enero-Mayo Para Tiempo Total



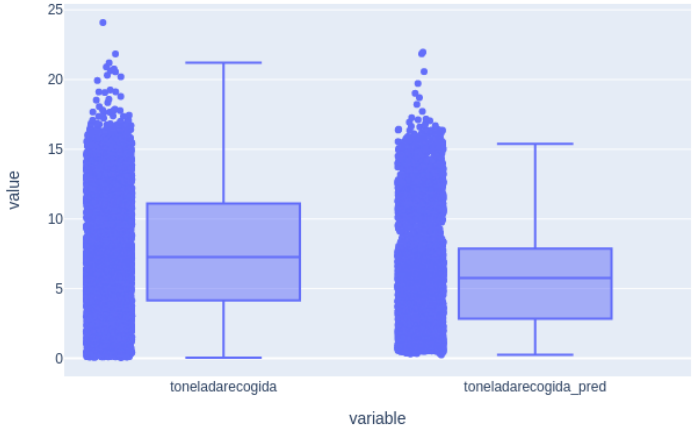
Elaboración Propia

Del mismo modo, en la Figura 37, se puede observar el resultado obtenido para la característica “Toneladas Recogidas”, el cual presenta mayores diferencias entre los datos reales y las predicciones realizadas por el modelo. Esto además se puede verificar con un valor R^2 de 0.4122.

Los resultados obtenidos en esta prueba demuestran cómo es indispensable mantener el reentrenamiento de los modelos en una franja de tiempo. Si bien el modelo para “Tiempo Total” no presentó una disminución considerable en la capacidad de ajuste a nuevos datos, las toneladas recogidas si tuvo esta disminución, esto debido a que el

comportamiento de recolección de basuras puede variar por circunstancias externas en el transcurso del tiempo.

Figura 37 Resultados Enero-Mayo Para Toneladas Recogidas



Elaboración Propia

Para finalizar, en la **fase 3,** se desarrolló un módulo de ciencia de datos para sofisticación de software Geoaseo en el cual se integró el producto en un prototipo donde se articula el cálculo de la capacidad del proceso resuelto en el objetivo 2, y los modelos en ciencia de datos en un software en un nivel tecnológico en TRL6,

A esta fase del proceso se le denominó “Desarrollo del módulo de ciencia de datos para sofisticación de software Geoaseo”.

Y principalmente se estableció llevar a producción los modelos de Tiempos Totales y Toneladas Recogidas que tuvieron un MSE menor y un R^2 con condiciones eficientes.

- **Análisis del sistema (ASI)**

Para este proceso se partió de un catálogo de requisitos, en el cual se identificaron las funcionalidades y objetos a integrar en el prototipo - Anexo_2_CATALOGO_REQUISITOS_GEOASEO_PRED

Seguidamente se establecieron los diccionarios de los datos para las variables a

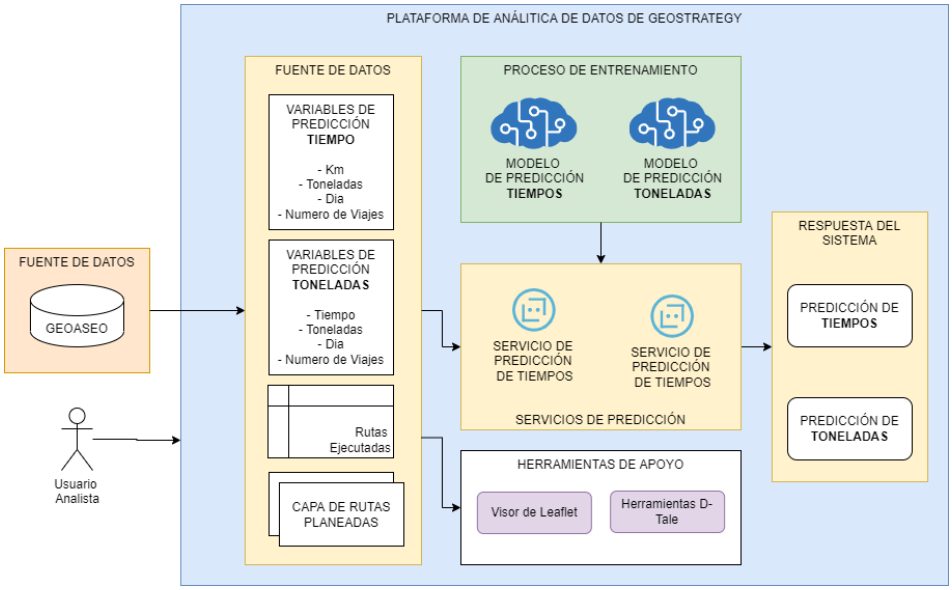
predecir.

Anexo_3_tab_dwh_t_prediccion_tiempo

Anexo_4_tab_dwh_t_prediccion_toneladas

Obteniendo así el siguiente modelo de negocio:

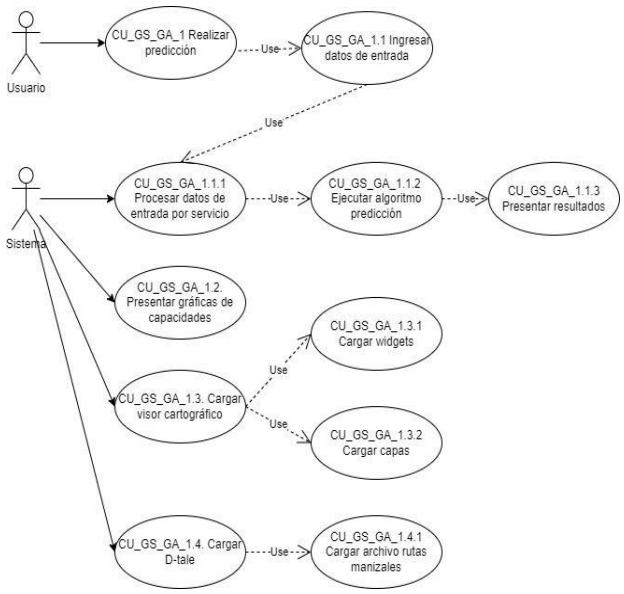
Figura 38 Modelo De Negocio Desarrollo De Software



Elaboración Propia

Y generando los siguientes casos de uso:

Figura 39 Casos de Uso Desarrollo De Software

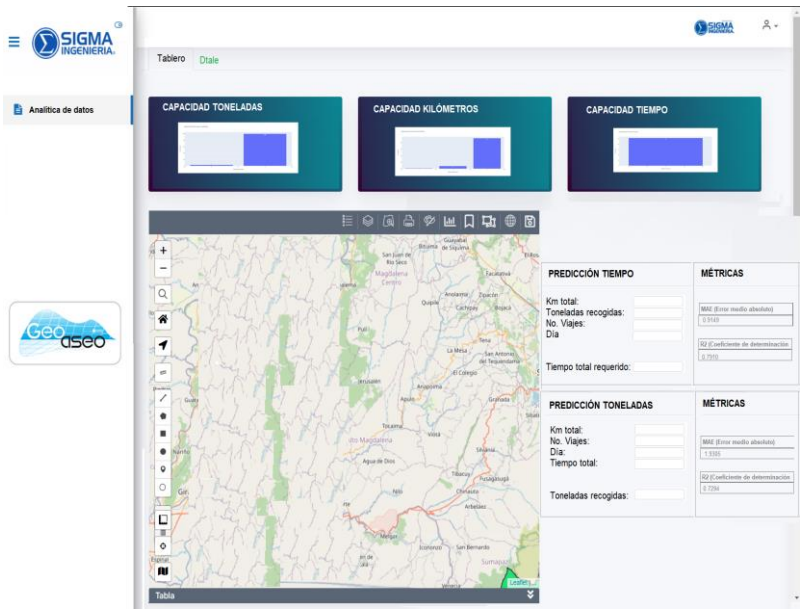


Elaboración Propia

De lo anterior, se pudo diseñar el funcionamiento del prototipo y generar la usabilidad el software, en el Anexo_5_DashBoard_Analitica_datos_Geoaseo se encuentra la descripción final del prototipo, con la usabilidad dada desde la lista de requerimiento.

A continuación, se presenta en la Figura 40, el mockup generado en la fase de análisis para llevar a diseño

Figura 40 Versión 1 Módulo de Ciencia De Datos Para Sofisticación De Software Geoaseo



Elaboración Propia

- **Diseño del sistema (DSI):** Integración en el sistema de los requerimientos especificados en el análisis de pruebas funcionales y no funcionales.

El desarrollo de la plataforma se compone de 2 partes esenciales, el frontend (Diseño) y el backend (Implementación), explicados a continuación:

Frontend.

Este corresponde a la parte del servicio con la que el usuario o cliente interactúa, buscando que el mismo cliente, en este caso el personal de la operación se sienta cómodo usando la plataforma. Para el caso específico de este proyecto el frontend está desarrollado en Angular

El diseño seleccionado para el menú de ingreso es el que se puede apreciar en la figura 41, dónde se tienen un módulo de validación de usuarios previamente creados.

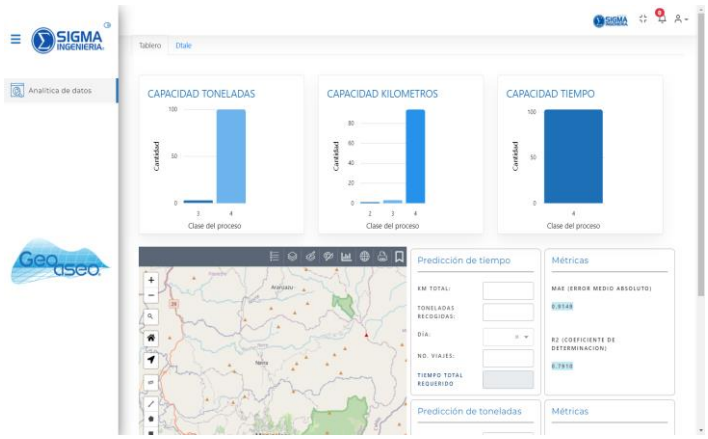
Figura 41 Menú de Ingreso De La Plataforma

Posterior al ingreso a la plataforma, se tendrá la pantalla principal. Como se muestra en la figura 42, en esta pantalla se tiene la posibilidad de ingresar los datos, para su posterior análisis, y así ver los resultados correspondientes. Cabe destacar que los nombres de las variables de ingresos son los siguientes:

- Predicción de tiempo: km total, toneladas recogidas, No de viajes y el día

- Predicción de toneladas: km total, tiempo total, No de viajes y el día

Figura 42 Pantalla Principal



En la zona de resultados se obtendrá un valor que indica los datos a recomendar por el modelo da para la operación de RSS correspondiente, mostrando además en una casilla en la parte lateral izquierda las métricas de evaluación, ya que en esta fase se lleva el proceso de validación en un entorno casi real.

Implantación (Prueba Piloto) (IAS): Implantación es el proceso de instalación para la disposición del cliente / usuario / piloto.

Backend

Esta es la parte con la que el usuario no interactúa, ya que en este se almacenan y organizan los datos, y también se asegura de que todo en el frontend de la plataforma funciona correctamente. Además, en esta se crean las librerías, y APIs para el procesamiento de los datos, teniendo en cuenta los códigos desarrollados en la fase 2, de desarrollo de los modelos basados en ciencia de datos. Para la puesta en marcha del servicio.

Inicialmente, es necesario realizar una validación entre desarrollo y ciencia de datos, de tal forma que se identifiquen las partes esenciales de los códigos a correr, así como los insumos propios de los modelos de inteligencia artificial que se deben cargar para realizar el procesamiento de los datos y retornar las salidas correspondientes.

En este orden de ideas, se identifican un total de 2 archivos, por cada una de las variables llevadas a predicción, siendo estos, archivos que representan los modelos de aprendizaje profundo. Posterior a la identificación de las partes clave, se debe crear un servicio en Django para la integración con Python. Django es un framework de desarrollo web de código abierto, escrito en Python, que respeta el patrón de diseño conocido como modelo–vista–controlador. De este modo, con Django se facilita el desarrollo del sitio web, ya que contiene una gran cantidad de funciones a la medida, y, además, posee una gran comunidad aportando, gracias a que es de código abierto (django, n.d.)

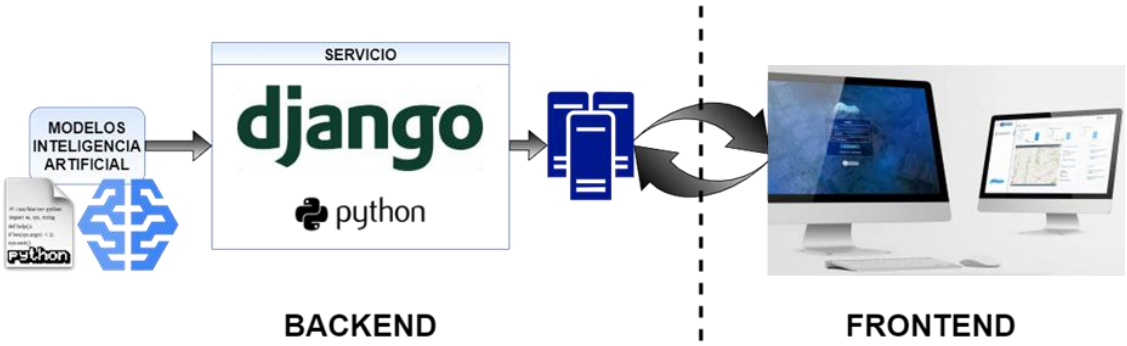
Luego que se ha montado el servicio, es necesario organizar un entorno virtual en Python, el cual contenga todos los requerimientos de librerías necesarios para el correcto funcionamiento de los códigos. Las librerías necesarias pertenecen todas a Python, en el Anexo_6_Librerias_servicio_prygeoaseo, se encuentran especificadas con su correspondiente versión

Posteriormente, para lograr un fácil y cómodo acceso por parte de los desarrolladores, en caso de que se deseen realizar futuros cambios a la plataforma, se deben organizar cada uno de los archivos siguiendo rutas en específico dentro del servidor. Del mismo modo, los códigos identificados por los desarrolladores deben organizarse como

funciones con entradas y salidas, que pueden ser llamadas desde cualquier parte del servidor.

Finalmente, se obtiene una URL de la API (Interfaz de Programación de Aplicaciones) que se conecta al frontend para el funcionamiento completo. Cabe resaltar que, para realizar todas las consultas y procesamiento de los datos, se hace en un clúster con las capacidades suficientes para cumplir esta tarea.

Figura 43 Estructura De La Plataforma



Elaboración Propia

En la Figura 43, se puede observar el procedimiento anteriormente descrito, desde la identificación de los modelos de inteligencia artificial, el montaje del servicio en django, bajo el lenguaje de programación Python, para finalmente ser usado en un clúster, el cual recibirá las peticiones de parte de los clientes (Frontend), realizará el procesamiento y retornará los resultados obtenidos a la plataforma.

En el siguiente enlace se encuentra el prototipo final con el respectivo usuario de ingreso

http://geostrategy.sigmaingenieria.net/

Login: Geoaseo

usuario: admin

contraseña: admin

En el Anexo_7_Manual de Usuario PRYGeoaseo, se encuentra un documento que da guía a la usabilidad del prototipo anteriormente descrito

9 DISCUSIÓN DE RESULTADOS

El primer elemento importante que se debe especificar en esta discusión es la importancia de los sistemas de información de captura de datos en las operaciones de servicios públicos domiciliarios de recolección y barrido, dichos sistemas de información en la medida que dispongan de una mayor automatización en captura de los datos pertinentes, aportan a la disminución de errores de información por captura manual, permitiendo hacia el futuro análisis retrospectivos mucho más específicos con un mayor nivel de calidad y de pertenencia en la operación.

Una vez los datos están recolectados correctamente, la extracción de la información es un procedimiento estratégico en la correcta interpretación del modelo, porque requiere de un entendimiento del sistema y del comportamiento de los datos desde el modelo de negocio.

Para el LSS la información es explicada en términos estadísticos y los comportamientos atípicos del modelo de negocios no son tenidos en cuenta en la mayoría de los casos, lo mismo puede ocurrir con la ciencia de datos que se olvidan las particularidades del modelo de negocio y los determina como “outlayer” o valor atípico del modelo sin interpretar los datos sin un previo conocimiento de los mismos. Es importante mantener un entendimiento del modelo de negocio y las particularidades que no son necesariamente explicadas por los datos.

Las ciencias de datos en la actualidad cuentan con herramientas que permiten identificar la carga significativa de las variables frente a un modelo y que permiten ratificar la voz del cliente, con respecto a sus variables críticas del modelo de negocio, modelos como PCA no solo se utilizan para la reducción de la dimensionalidad sino para determinar la importancia de variables dentro del modelo.

Si se entra en materia del modelo de caso en estudio se puede identificar la importancia de variables como kilómetros toneladas tiempos. Qué intuitivamente han sido estratégicas para la dirección de una operación de hacer, pero la ciencia de datos demuestra que en el caso de estudio conllevan el 51.30 de la variabilidad del modelo. No se puede descartar la importancia de variables como número de compactación, que

se excluyen desde este estudio porque en su captura de datos cuentan con vacíos en la información que no pudieron ser sustituidos a través de un mecanismo para ser incluidos dentro del modelo pero que en el análisis de componentes principales identifican un 24. 3% de la carga de variabilidad del modelo.

Por otro lado, la variable de combustible es una categoría de análisis qué sería conveniente dedicarle un trabajo especial para entender su comportamiento, sus límites de especificación y actualizarlo en la planeación operativa; está variable se omite desde esta investigación porque tiene una relación directa con los kilómetros recorridos, qué es uno de los elementos principales para la optimización y sobre los cuales el cliente puede tener control en el diseño. Sin embargo, como variable de impacto económico los combustibles pueden ser un elemento sensible dentro del modelo de negocio que amerita un análisis independiente, en una futura investigación.

La ciencia de datos contempla un área de conocimiento en las máquinas de aprendizaje, en el objeto de este estudio fue estratégico implementar máquinas de aprendizaje que fueran capaces de predecir las variables principales como el tiempo y las toneladas recogidas, este desarrollo en TRL6 pueden ser utilizados para predecir las nuevas planeaciones estratégicas y mejorar la capacidad del proceso o para ser utilizadas en línea y poder determinar en un momento específico de la operación cuál va a ser el comportamiento de la ruta a partir de las condiciones como el día, kilómetros, tiempo y número de viajes.

Como recomendación final de la investigación, es conveniente optimizar los parámetros de la planeación de las rutas para los límites de especificación de tiempo y toneladas, reflejando un comportamiento dinámico de las operaciones y manteniendo una capacidad del proceso acorde a las condiciones cambiantes del sistema, pero además estas técnicas de Machine Learning pueden ser utilizadas para calibrar el comportamiento de la ruta en línea.

Y finalmente, como logro de esta investigación se realizó un artículo científico tipo resumen, para el congreso ColCACI 2022, el cual fue aceptado para ponencia donde los trabajos aceptados (y presentados) se publicarán en las actas de la conferencia y estarán disponibles a través de **IEEE Xplore Digital Library® y** (según la política de IEEE) se

enviarán a la base de datos Scopus para su indexación en el Anexo_8_Articulo_GeoAseo_CD, se encuentra el documento enviado. Este proceso logro reducir la brecha entre la comunidad científica y la empresa porque permitió divulgar un proceso aplicado desde la investigación y desarrollo (ColCACI, 2022).

10 CONCLUSIONES

10.1 CONCLUSIONES OBJETIVO 1

La Etapa “Definir” dentro de LSS implica un entendimiento completo del modelo de negocio y sus características o variables que intervienen, además de establecer las especificaciones que el cliente requiere y entender cuáles son sus variables críticas dentro del modelo de negocio, es importante identificar el comportamiento de las variables que están comprometidas en el diseño de la operación.

Para la etapa de extracción, transformación y carga de los datos fue esencial contar con una plataforma tecnológica como GEOASEO que consolida la información, depura los datos, mantiene la histórica y minimiza el impacto de intervención humana; está última característica no se ve claramente sustentada en la variable de número de compactación que cuenta con una intervención humana y se evidencia en la distorsión los datos para esta variable.

Posteriormente, se inició el proceso de extracción de características sobre 200 rutas planeadas, y 59716 registros de ejecución, una vez depurados los datos desplegamos el análisis sobre 172 rutas, lo cual para empezar el entendimiento del modelo de negocio, fue muy importante la espacialización de la información porque permite tener un contexto general de lo que ocurre en la operación y un entendimiento de los datos alrededor de la planeación operativa. En la información estadística podemos observar qué es una operación geográficamente bien distribuida de forma equilibrada en cantidades de rutas y frecuencia con cobertura total sobre el territorio.

La dificultad de la planeación de la operación radica en el establecimiento de las metas y los límites de especificación para los elementos como toneladas planeadas, Km planeados, galones, número de compactaciones y número de viajes. Esta planeación operativa, en su diseño original no cuenta con la posibilidad de hacer uso de información histórica para determinar las metas y los límites de especificación de las variables críticas de los modelos de negocio.

10.2 CONCLUSIONES OBJETIVO 2

El objetivo de la fase medir es establecer la capacidad del proceso de las variables críticas del modelo de negocio, para el cliente en este caso, se identifica desde la etapa anterior la importancia de las toneladas, km recorridos, tiempos, número de compactaciones y combustible dentro del modelo de negocio.

En esta fase del proyecto entraron a medir 172 rutas que corresponden a 59716 registros de hojas de ruta ejecutadas, una vez aplicados las condiciones de limpieza de la información y de garantizar el conteo de las rutas que se han ejecutado más de 100 veces, el análisis de capacidad del proceso llevo a trabajar sobre 103 rutas. En primera instancia se determina por medio de un análisis de componentes principales, el nivel de importancia de las variables en el modelo entregando un 51.30% de variabilidad sobre las toneladas, tiempos y kilómetros. Esas tres variables confirman la expectativa del cliente en términos de controlar y gestionar dichas variables, pero además coinciden con los elementos a los cuales se les ha podido establecer los límites de especificación del proceso desde la extracción de los datos.

La siguiente etapa de la medición es establecer la capacidad para las variables que fueron seleccionadas a través del PCA, como toneladas, tiempo y kilómetros, para tal efecto se recorre las 103 rutas en cada una de sus ejecuciones calculando la capacidad del proceso y estableciendo como los LES y LEI son determinados en la planeación operativa obteniendo capacidades para cada una de las variables en estudio.

El número de compactación es una variable que tiene una gran carga de explicación del modelo de datos pero que cuenta con vacíos de información que no es posible sustituir luz desde la ejecución por tal motivo se recomienda un estudio posterior incluyendo está componente principal que tiene una carga del 24.95% de los datos, haciendo uso de técnicas de imputación de datos coherentes al modelo de negocio.

10.3 CONCLUSIONES OBJETIVO 3

En concordancia con los resultados presentados, las variables que permitieron una mejor predicción por métodos de regresión fueron “Tiempo Total" y “Toneladas Recogidas",

mostrando un mejor desempeño en todos los experimentos propuestos, evidenciado en las métricas de error usadas, principalmente en la regresión realizada con la red neuronal presentando valores de coeficiente de determinación R^2 de 0.79 y 0.73, respectivamente, además de los diagramas de caja. Del mismo modo, el modelo que demostró el mejor desempeño fue el de redes neuronales totalmente conectadas, logrando el menor error en la regresión para cada una de las variables en todos los experimentos, resultado que puede estar atribuido a la complejidad que estás presentan y la mayor cantidad de operaciones realizadas en comparación a los otros métodos evaluados. Asimismo, teniendo en cuenta la naturaleza del experimento 3 se puede evidenciar cómo la estandarización de los datos, aunque no representan un factor determinante para la reducción de los porcentajes de error, presenta beneficios reflejados en la reducción del error, esto debido a la gran variabilidad de rangos presentes en cada una de las variables de entrada a los modelos.

Los resultados presentados permiten la predicción de variables como “Tiempo Total" y “Toneladas Recogidas” fundamentales en el proceso de recolección de residuos sólidos urbanos, y que pueden ser optimizadas al realizar una predicción basada en datos reales, factor que permitiría la optimización de tiempos y beneficios para el proceso en general.

Como trabajo futuro, se propone la implementación de los modelos desarrollados, cómo factores de prueba de los procesos, además de un acompañamiento en la recolección de una mayor cantidad de datos, que permitan optimizar los modelos implementados, así como la evaluación en otras zonas de estos. Adicionalmente, se propone el estudio y optimización de más algoritmos de regresión y de preprocesamiento de los datos, ya que estos pueden ser representativos en las tareas de regresión.

11 RECOMENDACIONES

Se recomienda un estudio posterior, por ejemplo, de importancia relativa, más específico de la variable número de compactaciones para entender la relación que existe con las rutas y por qué esta tiene un mayor peso en la componente principal # 2 la cual explica el 24.95% de la variabilidad de los datos.

Se propone un método de imputación de datos para los valores perdidos en la variable número de compactaciones con lo cual se podría estudiar su importancia dentro de los objetivos de mejora.

Seguir trabajando en la automatización y captura automática de los datos desde la fuente de origen es una garantía para las empresas de servicios públicos de poder contar con datos óptima para hacer analítica de datos predicción Machine Learning y trabajar permanentemente en la optimización de procesos de una forma continua y controlada.

Cómo trabajo futuro el análisis de causa raíz puede ser una investigación de carácter doctoral qué implica crear el modelo de causalidad con las diferentes internas o externas que intervienen en el modelo, en búsqueda de identificar los puntos de optimización del modelo de recolección de residuos sólidos urbanos, asistida por Machine Learning.

Las máquinas de aprendizaje diseñadas para esta investigación pueden ser el principio de una investigación doctoral para hacer diseño automático de rutas a partir de las predicciones de kilómetros y toneladas, en el modelo de negocio lograr articular predicción de estos, pueden establecer los requerimientos técnicos requeridos para diseñar rutas automáticas que cumplan y satisfagan este conjunto de variables generadas por predicción de los modelos.

Sí bien es cierto que la red neuronal completamente conectada que logra hacer la predicción de las diferentes variables ha logrado muy buen desempeño para este modelo es conveniente como un trabajo futuro experimentar diferentes tipos de redes neuronales y diferentes configuraciones de nodos y de capas que permitan seguir

profundizando en los posibles ajustes a los hiper-parámetros que logren un mejor desempeño del modelo.

12 REFERENCIAS

Adelheid Januaviani, T. M., Gusriani, N., & Joebaedi, K. (2019). The Best Model of LASSO With The LARS (Least Angle Regression and Shrinkage) Algorithm Using Mallow's Cp. http://www.worldscientificnews.com/article-in-press/2019-2/115-118-2019/

Aguilar Pirachicán, C. M. (2018). Propuesta de un marco general para el despliegue de ciudades inteligentes apoyado en el desarrollo de IoT en Colombia. https://repository.usta.edu.co/handle/11634/10732

Albañil, E. J., & Martínez, K. N. (2019). Propuesta para implementar Lean Six Sigma en el departamento de servicio al cliente en una empresa del sector retail. Ingeniería Industrial. https://ciencia.lasalle.edu.co/ing_industrial/108

Alvarado Chávez, F. B. (2018). Mejora de procesos ERP´S (enterprise resource planning) con lean six sigma - Dialnet. https://dialnet.unirioja.es/servlet/articulo?codigo=6839148

Aouag, H., & Mohyiddine, S. (2023). Improvement of Lean Manufacturing approach based on MCDM techniques for sustainable manufacturing. International Journal of Manufacturing Research, 18(1), 1. https://doi.org/10.1504/IJMR.2023.10040708

Aquino, A. A. (2015). Hacia un nuevo proceso de minería de datos centrado en el usuario. 680.

Araiza Aguilar, J. A., & José Zambrano, M. E. (2015). Mejora del servicio de recolección de residuos sólidos urbanos empleando herramientas SIG: un caso de estudio. https://www.redalyc.org/pdf/467/46750925005.pdf

Arencibia-Jorge, R., Leydesdorff, L., Chinchilla-Rodríguez, Z., Rousseau, R., & Paris, S. W. (2009). Retrieval of very large numbers of items in the Web of Science: an exercise to develop accurate search strategies. El Profesional de la Información, 18(5), 529-533.

Arcidiacono, G., Costantino, N., & Yang, K. (2016). The AMSE Lean Six Sigma governance model. International Journal of Lean Six Sigma, 7(3), 233–266. https://doi.org/10.1108/IJLSS-06-2015-0026/REFERENCES

Arcidiacono, G., & Pieroni, A. (2018). The Revolution Lean Six Sigma 4.0. International Journal on Advanced Science, Engineering and Information Technology, 8(1), 141–149. https://doi.org/10.18517/IJASEIT.8.1.4593

Akbarpour, N., Salehi-Amiri, A., Hajiaghaei-Keshteli, M., & Oliva, D. (2021). An innovative waste management system in a smart city under stochastic optimization using vehicle routing problem. Soft Computing, 25(8), 6707-6727. https://doi.org/10.1007/s00500-021-05669-6

Ayming. (n.d.). ¿Qué son los TRL (Technology Readiness Levels)? - Ayming España. Retrieved May 30, 2022, from https://www.ayming.es/insights-y-noticias/noticias/trl-technology-readiness-levels/

Bedoui, A., & Lazar, N. A. (2020). Bayesian empirical likelihood for ridge and lasso regressions. Computational Statistics & Data Analysis, 145, 106917. https://doi.org/10.1016/J.CSDA.2020.106917

Belyadi, H., & Haghighat, A. (2021). Machine Learning Guide for Oil and Gas Using Python (pp. 169-295). Gulf Professional Publishing.

Beall, J. (2011). Academic Library Databases and the Problem of Word-Sense Ambiguity. The Journal of Academic Librarianship, 37(1), 64-69. doi:10.1016/j.acalib.2010.10.008

Barber, E., Pisano, S., Romagnoli, S., Parsiale, V., De Pedro, G., & Gregui, C. (2008). Los catálogos en línea de acceso público del Mercosur disponibles en entorno web. Información, Cultura y Sociedad, (18), 37-55.

Bento da Silva, I., Godinho Filho, M., Agostinho, O., & Lima, O. (2019). A new Lean Six Sigma framework for improving competitiveness. https://www.redalyc.org/journal/3032/303260200027/303260200027.pdf

Betanzo-Quezada, Eduardo, Torres-Gurrola, Miguel Ángel, Romero-Navarrete, José Antonio, & Obregón-Biosca, Saúl Antonio. (2016). Evaluación de rutas de recolección de residuos sólidos urbanos con apoyo de dispositivos de rastreo satelital: análisis e implicaciones. Revista internacional de contaminación ambiental, 32(3), 323-337. https://doi.org/10.20937/RICA.2016.32.03.07

Calvo Mazuera, S. E. (2020). Estadística Descriptiva: Conceptos y Visualizaciones . https://books.google.com.co/books?id=6ezczQEACAAJ&dq=Estad%C3%ADstica+Descriptiva+2020&hl=es&sa=X&redir_esc=y

CEPAL. (2020). Gasto público para impulsar el desarrollo económico e inclusivo y lograr los Objetivos de Desarrollo Sostenible. https://repositorio.cepal.org/bitstream/handle/11362/46276/1/S2000670_es.pdf

Cheng, C. (2022). Principal Component Analysis (PCA) Explained Visually with Zero Math | by Casey Cheng | Towards Data Science. https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

ColCACI. (2022). Inscripción - ColCACI 2022. https://ieee-colcaci.org/2022/registration/

Copaja Alegre, M., & Esponda Alva, C. (2019). Tecnología e innovación hacia la ciudad inteligente. Avances, perspectivas y desafíos. Bitácora Urbano Territorial, 29(2), 59–70. https://doi.org/10.15446/bitacora.v29n2.68333

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online Passive-Aggressive Algorithms. Journal Of Machine Learning Research, 7, 551-585. Retrieved 23 May 2022.

Dedić, N., & Stanier, C. (2017). Towards differentiating business intelligence, big data, data analytics and knowledge discovery. Lecture Notes in Business Information Processing, 285, 114–122. https://doi.org/10.1007/978-3-319-58801-8_10/FIGURES/1

django. (n.d.). Django REST framework. Retrieved May 30, 2022, from https://www.django-rest-framework.org/

DNP. (2016). POLÍTICA NACIONAL PARA LA GESTIÓN INTEGRAL DE RESIDUOS SÓLIDOS.

DNP, D. N. de P. (2008). LINEAMIENTOS Y ESTRATEGIAS PARA FORTALECER EL SERVICIO PÚBLICO DE ASEO EN EL MARCO DE LA GESTIÓN INTEGRAL DE RESIDUOS SÓLIDOS. CONPES 3530. https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3530.pdf

Doga, O., & Faruk Gurcan, O. (2018). Data Perspective of Lean Six Sigma in Industry 4.0 Era: A Guide To Improve Quality. http://www.ieomsociety.org/paris2018/papers/170.pdf

EPA. (2022). Managing and Reducing Wastes: A Guide for Commercial Buildings | US EPA. https://www.epa.gov/smm/managing-and-reducing-wastes-guide-commercial-buildings

Evalue. (2020). TRL o los niveles de madurez tecnológica - Evalue Innovación. https://www.evalueconsultores.com/los-niveles-de-madurez-tecnologica-trl/

Felizzola Jiménez, H., & Luna Amaya, C. (2014). Lean Six Sigma en pequeñas y medianas empresas: un enfoque metodológico. Revista Chilena de Ingeniería, 22(2), 263–277. http://www.redalyc.org/articulo.oa?id=77231016012

Flores Lagla, G. A., Cadena Moreano, J. A., Quinatoa Arequipa, E. E., & Villa Quishpe, M. W. (2019). Minería de datos como herramienta estratégica - Dialnet. https://dialnet.unirioja.es/servlet/articulo?codigo=6796766

Garza Ríos, R. C., Hernández, C. M., Rodríguez, G. E. L., & González Sánchez, C. N. (2016). Aplicación de la metodología DMAIC de Seis Sigma con simulación discreta y técnicas multicriterio. Revista de Métodos Cuantitativos Para La Economía y La Empresa. https://www.redalyc.org/articulo.oa?id=233148815002

Garzón, N. A., González Neira, E. M., & Pérez Vélez, I. (2017). Metaheurística para la solución del Problema de diseño de la red de transporte multiobjetivo con demanda multiperiodo. Ingeniería y Ciencia, 13(25), 29–69. https://doi.org/10.17230/INGCIENCIA.13.25.2

Góngora, G. P. M., & Bernal, W. N. (2015). Factores Clave en la Gestión de Tecnología de Información para Sistemas de Gobierno Inteligente. Journal of Technology Management & Innovation, 10(4), 109–117. https://doi.org/10.4067/S0718-27242015000400012

Grech, V., & Calleja, N. (2018). WASP (Write a Scientific Paper): Multivariate analysis. Early Human Development, 123, 42–45. https://doi.org/10.1016/J.EARLHUMDEV.2018.04.012

Greener, R. (2020). Stop testing for normality. Normality tests are misleading and a… | by Robert Greener | Towards Data Science. https://towardsdatascience.com/stop-testing-for-normality-dba96bb73f90

Gutiérrez Pulido, H., & de la Vara Salazar, R. (2013). Control estadístico de la calidad y Seis Sigma. McGraw-Hill Interamericana.

Gutiérrez Pulido, H., & de la Vara Salazar, R. (2009). CONTROL ESTADÍSTICO DE CALIDAD Y SEIS SIGMA (Segunda edición). https://www.uv.mx/personal/ermeneses/files/2018/05/6-control-estadistico-de-la-calidad-y-seis-sigma-gutierrez-2da.pdf

Hannan, M., Hossain Lipu, M., Akhtar, M., Begum, R., Al Mamun, M., & Hussain, A. et al. (2020). Solid waste collection optimization objectives, constraints, modeling approaches, and their challenges toward achieving sustainable development goals. Journal Of Cleaner Production, 277, 123557. https://doi.org/10.1016/j.jclepro.2020.123557

Herrera Vidal, G., Pérez Aguas, Y., & Venecia Puello, E. (2017). Enfoque seis sigma y proceso analítico jerárquico en empresa del sector lácteo 1. https://www.redalyc.org/articulo.oa?

Herrera, F., Herrera-Viedma, E., Alonso, S., & Cabrerizo, F.J. (2009). Agregación de índices bibliométricos para evaluar la producción científica de los investigadores. El Profesional de la Información, 18(5), 559-561.

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. PLoS Computational Biology, 4(10), e1000204. doi:10.1371/journal.pcbi.1000204

IEEE. (2014). Artificial neural network-based time series analysis forecasting for the amount of solid waste in Bangkok. In Ninth International Conference on Digital Information Management (ICDIM 2014) (pp. 16-20). Phitsanulok, Thailand. Retrieved 23 May 2022, from http://10.1109/ICDIM.2014.6991427

IBM. (2022). Ventajas de la programación de restricciones - Documentación de IBM. https://www.ibm.com/docs/es/icos/20.1.0?topic=programming-benefits-constraint

Icarte Ahumada, G. A. (2016). Aplicaciones de inteligencia artificial en procesos de cadenas de suministros: una revisión sistemática. Ingeniare. Revista Chilena de Ingeniería, 24(4), 663–679. https://doi.org/10.4067/S0718-33052016000400011

Jammeli, H., Ksantini, R., Ben Abdelaziz, F., & Masri, H. (2021). Sequential Artificial Intelligence Models to Forecast Urban Solid Waste in the City of Sousse, Tunisia. IEEE Transactions On Engineering Management, 1-11. https://doi.org/10.1109/tem.2021.3081609

Martínez, R. (2007). Biblioteca Digital: conceptos, recursos y estándares. Buenos Aires: Alfagrama.

Medina Rojas, F., & Gámez Santamaria. Cristina. (2014). Funcionalidades de la minería de datos | Ingeniería y Región. https://journalusco.edu.co/index.php/iregion/article/view/728/1395

M.Howari, F., & Ghrefat, H. (2021). Geographic information system: spatial data structures, models, and case studies. In Pollution Assessment for Sustainable

Practices in Applied Sciences and Engineering (pp. 165-198). Retrieved 23 May 2022, from https://doi.org/10.1016/C2015-0-06451-6.

MINVIVIENDA. (2013). Decreto 2981 de 2013 - Gestor Normativo - Función Pública. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56035

Montalvo, F. D. (2018). Brecha del servicio de limpieza pública en la ciudad de Tingo Maria, Perú. Anales Científicos, 79(2), 291–297. https://doi.org/10.21704/AC.V79I2.1241

Muñoz, W. Z. (2018). Proyectos de desarrollo de proveedores que usan Six Sigma: un análisis de caso en Schneider Electric Colombia S.A. Revista Escuela de Administración de Negocios, 173–184. https://doi.org/10.21158/01208160.N0.2018.2023

NASA. (2012). Technology Readiness Level | NASA. https://www.nasa.gov/directorates/heo/scan/engineering/technology/technology_readiness_level

Ogryzek, M., & Wolny-Kucí, A. (2021). Geo-Information Sustainable Development of Transport as a Regional Policy Target for Sustainable Development-A Case Study of Poland. https://doi.org/10.3390/ijgi10030132

PAe - Métrica v.3. (n.d.). Retrieved May 30, 2022, from https://administracionelectronica.gob.es/pae_Home/pae_Documentacion/pae_Metodolog/pae_Metrica_v3.html#.YpaVOyjMK3A

Passive Aggressive Algorithm—For big data models. Medium. (2021). Retrieved 23 May 2022, from https://medium.com/geekculture/passive-aggressive-algorithm-for-big-data-models-8cd535ceb2e6.

Palacios, H. J. G., Toledo, R. A. J., Pantoja, G. A. H., & Navarro, Á. A. M. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. Advances in Science,

Technology and Engineering Systems, 2(3), 598–604. https://doi.org/10.25046/AJ020376

Patiño Chirva, J. A., Daza Cruz, Y. X., & López-Santana, E. R. (2016). Un Enfoque Híbrido de Agrupamiento y Optimización Entera Mixta para el Problema de Servicios de Recoleccion Selectiva de Residuos Sólidos Domésticos. Ingeniería, 21(2), 235–257. https://doi.org/10.14483/UDISTRITAL.JOUR.REVING.2016.2.A09

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830

PostgreSQL. (n.d.). PostgreSQL: The world’s most advanced open source database. Retrieved June 1, 2022, from https://www.postgresql.org/

Python. (n.d.). Welcome to Python.org. Retrieved June 1, 2022, from https://www.python.org/

QGIS. (n.d.). Bienvenido al proyecto QGIS! Retrieved June 1, 2022, from https://qgis.org/es/site/

Ramsundar, B., & Zadeh, R. (2018). TensorFlow for Deep Learning: From Linear Regression to Reinforcement Learning. O'Reilly Media, Inc

RapidMiner. (n.d.). RapidMiner | Amplify the Impact of Your People, Expertise & Data. Retrieved May 31, 2022, from https://rapidminer.com/

Rivera Cerpa, Y. M., & Conrado Tobón, J. (2016). Impacto de los líderes en la productividad de las empresas de servicio de aseo en la ciudad de Barranquilla - Dialnet. https://dialnet.unirioja.es/servlet/articulo?codigo=6104144

Rojas Salazar, M. L., & Pérez Olguín, I. J. C. (2019). Ciclo DMAIC en Latinoamérica: Análisis de aplicación y relación con el Producto Interno Bruto. https://www.researchgate.net/publication/333077156_Ciclo_DMAIC_en_Latinoamerica_Analisis_de_aplicacion_y_relacion_con_el_Producto_Interno_Bruto

Rudolph Raj, J., & Seetharaman, A. (2013). Role of waste and performance management in the construction industry. Journal of Environmental Science and Technology, 6(3), 119–129. https://doi.org/10.3923/JEST.2013.119.129

Sáez, A. (2011). Factores críticos para la medición de la calidad de servicio del aseo urbano en el municipio Maracaibo* Critical Factors for Measuring the Quality of Urban Sanitary Services in the Maracaibo Municipality. 13(3), 329–354.

Sayad, S. (2022). Support Vector Regression. Saedsayad.com. Retrieved 23 May 2022, from https://www.saedsayad.com/support_vector_machine_reg.htm.

Sánchez-Muñoz, M. del P., Cruz-Cerón, J. G., & Maldonado-Espinel, P. C. (2020). Urban solid waste management in Latin America: An analysis from the perspective of waste generation. Revista Finanzas y Politica Economica, 11(2), 321–336. https://doi.org/10.14718/REVFINANZPOLITECON.2019.11.2.6

Sarria Yépez, M. P., Fonseca Villamarín, G. A., & Bocanegra Herrera, C. C. (2017). Modelo metodológico de implementación de lean manufacturing. https://doi.org/10.21158/01208160.n83.2017.1825

Scikit Learn. (2022a). sklearn.linear_model.LassoLars — scikit-learn 1.1.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLars.html

Scikit Learn. (2022b). Standard Scaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Serrano, E. V. (2016). Estudio exploratorio en torno a las potencialidades de los recicladores de oficio para la construcción de nueva política pública con inclusión social en el sistema de aseo en Bogotá D. C. (Colombia). Reflexión Política, 18(35), 98–113. https://doi.org/10.29375/01240781.2467

SIGMA INGENIERIA S.A. (n.d.). Software de georeferenciacion en Colombia, Sigma Ingeniería, expertos en SIG. Retrieved May 30, 2022, from https://www.sigmaingenieria.com.co/#geoambiental

Sra, S., Nowozin, S., & Wright, S. (2011). Optimization for Machine Learning. The MITPress.

Sodanil, M., & Chatthong, P. (2014). Artificial neural network-based time series analysis forecasting for the amount of solid waste in Bangkok. 2014 9th International Conference on Digital Information Management, ICDIM 2014, 16–20. https://doi.org/10.1109/ICDIM.2014.6991427

Solano, J. K. (2021). Propuesta metodológica basada en redes neuronales artificiales para la determinación de la gestión óptima de residuos sólidos urbanos: aplicación en las localidades de Suba y Engativá de la ciudad de Bogotá (Colombia). Universidad de Valencia. https://riunet.upv.es/bitstream/handle/10251/168119/Solano%20-%20Propuesta%20metodol%C3%B3gca%20basada%20en%20redes%20neuronales%20artfcales%20para%20la%20determnac%C3%B3n%20de%20la%20ges....pdf?sequence=1

Soto Mejía, J., Solarte Martínez, G. R., & Muñoz Guerrero, L. E. (2019). Localización del punto óptimo de partida en el problema de ruteo vehicular con capacidad restringida (CVRP). Tecnura, 23(59), 27–46. https://doi.org/10.14483/22487638.13653

Superservicios. (2019). Disposición Final de Residuos Sólidos Informe Nacional– 2018. Superintendencia de Servicios Públicos Domiciliarios. https://www.superservicios.gov.co/sites/default/archivos/Publicaciones/Publicaciones/2020/Ene/informe_nacional_disposicion_final_2019_1.pdf

Tampubolon, S., & Purba, H. H. (2021). Lean six sigma implementation, a systematic literature review. International Journal of Production Management and Engineering, 9(2), 125–139. https://doi.org/10.4995/IJPME.2021.14561

Tarín, A. R. (2018). El paradigma de las Smart Cities en el marco de la gobernanza urbana. Gestión y Análisis de Políticas Públicas, 29–35. https://doi.org/10.24965/GAPP.V0I20.10536

Trujillo González, J. M., Niño Torres, Á. M., & Niño Torres, A. P. (2017). Gestión de residuos sólidos domiciliarios en la ciudad de Villavicencio. Una mirada desde los grupos de interés: empresa, estado y comunidad. https://www.redalyc.org/articulo.oa?id=321750362011

Vargas-Hernández, J. G., Castillo, M. T. J., & Muratalla-Bautista, G. (2018). Sistemas de producción competitivos mediante la implementación de la herramienta Lean Manufacturing. Ciencias Administrativas, 11, 020–020. https://doi.org/10.24215/23143738E020

Wink, D. M., & Killingsworth, E. K. (2011). Optimizing use of library technology. Nurse Educator, 36(2), 48-51.

Yang, X.-S. (2021). Genetic Algorithms. Nature-Inspired Optimization Algorithms, 91–100. https://doi.org/10.1016/B978-0-12-821986-7.00013-5

Zhang, H., Ge, H., Yang, J., & Tong, Y. (2021). Review of Vehicle Routing Problems: Models, Classification and Solving Algorithms. Archives Of Computational Methods In Engineering, 29(1), 195-221. https://doi.org/10.1007/s11831-021-09574-x