ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Estimación de tiempos de carga de transportes mediante modelos Machine Learning para una planta de distribución de productos de nutrición de cultivo

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

Galo Alexander Reyes Clemente

Kevin Franklin Sisalima Jiménez

GUAYAQUIL - ECUADOR

Año: 2025

DEDICATORIA

El presente proyecto lo dedico, en primer lugar, a mí mismo, por haber tenido el coraje de salir de mi zona de confort y atreverme a explorar un mundo nuevo: el de los datos. Este camino no ha sido fácil, pero me ha enseñado a pensar diferente, a encontrar sentido en la complejidad y a confiar en mis propias capacidades.

A mi familia, por su amor incondicional, su apoyo constante y por ser siempre mi refugio y mi motivación. Sin su presencia, esta meta no habría tenido el mismo valor.

A mis profesores, quienes con paciencia, exigencia y conocimiento supieron guiarme en este proceso. Gracias por compartir su experiencia y sembrar en mí la pasión por aprender.

Esta dedicatoria es un reconocimiento a cada paso dado, a cada error convertido en lección, y a cada persona que, directa o indirectamente, hizo posible este logro.

AGRADECIMIENTOS

Mi más sincero agradecimiento a la ESPOL y a FIEC, por brindarnos una formación académica integral, desafiante y de alta calidad.

Extiendo un agradecimiento especial al profesor Danny Alfredo Torres Moran, por su guía comprometida, sus valiosos aportes técnicos y su acompañamiento constante durante el desarrollo de este proyecto.

Agradezco profundamente a la empresa de distribución de productos de nutrición de cultivos, por permitirnos acceder a su información operativa, confiar en nuestro trabajo y mostrarse abierta a la innovación basada en ciencia de datos. Su colaboración fue fundamental para la ejecución y relevancia de esta investigación.

Finalmente, a mi familia, por ser mi motor en todo momento. Gracias por su apoyo incondicional, sus palabras de aliento y su paciencia, que me acompañaron en cada etapa de este proceso.

A todos ustedes, gracias por contribuir a la realización de este logro académico y personal.

DECLARACIÓN EXPRESA

Nosotros Kevin Franklin Sisalima Jiménez y Galo Alexander Reyes Clemente acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique a los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 22 de diciembre del 2025

Firmado electrónicamente por:
GALO ALEXANDER
REYES CLEMENTE

Galo Alexander
Reyes Clemente

Firmado electrónicamente por:
KEVIN FRANKLIN
SISALIMA JIMENEZ
Validar únicamente con FirmaEC

Kevin Franklin
Sisalima Jiménez

COMITÉ EVALUADOR

Firmado electrónicamente por:
DANNY ALFREDO TORRES MORAN
Validar únicamente con FirmaEC

Maria Isabel Mera Collantes
Time Stamping Security Data

Danny Alfredo Torres Moran

PROFESOR TUTOR

Maria Isabel Mera Collantes

PROFESOR EVALUADOR

RESUMEN

El presente proyecto tiene como objetivo implementar una herramienta de visualización interactiva basada en modelos de aprendizaje automático para apoyar la planificación logística y mejorar la eficiencia del proceso de carga en una planta de productos de nutrición de cultivos. Se aplicó la metodología CRISP–DM para estructurar el análisis, incluyendo la comprensión del negocio, el análisis exploratorio, la preparación de datos, el modelado predictivo y el despliegue en Power BI.

Se entrenaron modelos de regresión supervisada (Regresión Lineal, Random Forest y XGBoost), determinándose que Random Forest ofrece el mejor desempeño al reducir el error promedio de planificación de 14.02 a 7.38 minutos, lo que representa una mejora del 47.3 % respecto al método manual utilizado por la planta. La herramienta desarrollada permite visualizar estimaciones de tiempo de carga por jornada, comparar valores reales y predichos, e identificar desviaciones operativas mediante indicadores como OTIF.

El análisis económico evidencia un ahorro operativo estimado entre USD 1,200 y USD 1,600 mensuales, equivalente a aproximadamente USD 18,000 anuales, sin considerar beneficios indirectos como la reducción de tiempos muertos y la mayor estabilidad del proceso. El estudio demuestra la viabilidad de integrar técnicas de machine learning y herramientas de inteligencia de negocios en entornos logísticos reales, aportando una base metodológica aplicable a otras operaciones del sector.

ABSTRACT

This project aims to implement an interactive visualization tool based on machine learning models to support logistical planning and improve loading efficiency in a crop nutrition production plant. The CRISP–DM methodology was applied to structure the process, including business understanding, exploratory analysis, data preparation, predictive modeling, and deployment within Power BI.

Supervised regression models (Linear Regression, Random Forest, and XGBoost) were trained, with Random Forest achieving the best performance. The model reduced the average planning error from 14.02 to 7.38 minutes, representing a 47.3% improvement compared to the current manual estimation method used by the plant. The visualization tool enables planners to view predicted loading times, compare them with actual values, and identify operational deviations through indicators such as OTIF.

The economic assessment shows operational savings of approximately USD 1,200 to USD 1,600 per month, equivalent to around USD 18,000 annually, excluding indirect benefits such as reduced idle time and improved process stability. The study demonstrates the feasibility of integrating machine learning techniques with business intelligence tools in real logistics environments, providing a methodological framework that can be replicated in similar industrial operations.

Keywords: logistics, machine learning, Random Forest, CRISP–DM, data visualization, Power BI, loading time estimation.

ÍNDICE GENERAL

ABREVIATURAS

ESPOL: Escuela Superior Politécnica del Litoral

ML: Machine Learning

IA: Inteligencia Artificial

ML: Machine Learning

RF: Random Forest

XGB: XGBoost

PCA: Principal Component Analysis

CRISP-DM: Cross-Industry Standard Process for Data Mining

RMSE: Root Mean Squared Error

MAE: Mean Absolute Error

R^2: Coeficiente de determinación

OTIF: On Time In Full

SQL: Structured Query Language

SAP: Systems, Applications and Products in Data Processing.

Min: minutos

Hh: hora-hombre

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

CAPÍTULO 1

1.1. DESCRIPCIÓN DEL PROBLEMA

La eficiencia operativa del proceso de carga en plantas de distribución de productos de nutrición de cultivos es esencial para asegurar el cumplimiento de cronogramas y la satisfacción del cliente. No obstante, la gestión actual se realiza sin el respaldo de herramientas analíticas que capitalicen los datos disponibles para anticipar desviaciones o mejorar la planificación (Singh, 2021; Zhou & Wang, 2019).

Esta carencia se traduce en diversas limitaciones operativas:

- Alta variabilidad en los tiempos de carga, incluso bajo condiciones similares.
- Cuellos de botella recurrentes que ralentizan el proceso.
- Dificultades en la asignación óptima de recursos humanos y físicos.
- Indicadores de eficiencia poco precisos debido al uso de tiempos planificados con criterios empíricos (Araujo & Etemad, 2020).

El uso del indicador "sacos por hora-hombre" basado en tiempos asignados sin análisis formal genera una holgura artificial que distorsiona la percepción del rendimiento real. En términos Lean, esto constituye un desperdicio que impide detectar ineficiencias estructurales y limita el avance hacia la mejora continua (Ohno, 1988; Womack & Jones, 2003).

Aunque se cuenta con registros históricos de las operaciones logísticas, estos no se han utilizado de forma sistemática con herramientas avanzadas como el machine learning. La literatura evidencia que el uso de modelos predictivos permite estimar con mayor precisión los tiempos de carga y mejorar la toma de decisiones operativas en entornos logísticos complejos (Zhou & Wang, 2019; Singh, 2021; Araujo & Etemad, 2020).

La falta de una herramienta visual interactiva también limita la capacidad de supervisión en tiempo real y la reacción ante eventualidades, derivando en una gestión reactiva en lugar de proactiva. Estudios recientes demuestran que la integración de modelos predictivos con visualización dinámica mejora la eficiencia y reduce la incertidumbre operativa (Singh, 2021; Zhou & Wang, 2019; Araujo & Etemad, 2020).

En la Tabla 1, se evidencia que en julio de 2024 existe una diferencia de hasta 25 minutos en los tiempos reales de carga de transportes. El transporte 80006243, con un tiempo planificado de 40 minutos, completó su carga en tan solo 15 minutos; mientras que el

transporte 80006323, con 60 minutos planificados, requirió 36 minutos reales. Esta variabilidad de hasta 25 minutos revela cómo la planificación basada en estimaciones no estandarizadas puede introducir distorsiones importantes en la percepción de eficiencia operativa.

TRANSPORTE	FECHA INICIO DE CARGA	TIEMPO PLANIFICADO DE CARGA	TIEMPO REAL DE CARGA	PESO A CARGAR	PRESENTACION	SACOS A CARGAR	LINEA DE PRODUCCIÓN
80006243	3/7/2024	40 MIN	15 MIN	9500	50 KG	190	2
80006323	5/7/2024	60 MIN	36 MIN	25000	50 KG	500	2

Tabla 1 Ejemplos de variabilidad del tiempo de carga planificado vs real

El presente estudio aborda estas brechas mediante el desarrollo de una solución basada en datos, orientada a fortalecer la planificación logística y consolidar una operación más ágil, confiable y orientada al cliente.

1.2. JUSTIFICACIÓN DEL PROBLEMA

El uso de modelos predictivos en logística representa una estrategia eficaz para optimizar la planificación, reducir tiempos improductivos y asignar recursos de manera inteligente (Zhou & Wang, 2019; Singh, 2021). En contextos como el de la planta de distribución analizada, donde la variabilidad y la planificación no estandarizada son predominantes, una solución basada en evidencia resulta crítica para mejorar el desempeño operativo (Araujo & Etemad, 2020).

Actualmente, la eficiencia operativa se mide mediante el indicador “sacos por hora-hombre”, el cual depende directamente del tiempo estimado por el área de planificación. Cuando estos tiempos se asignan sin un sustento analítico adecuado, los resultados pueden no representar la realidad del proceso. Esta situación compromete la evaluación objetiva del desempeño logístico y limita las posibilidades de mejora continua (Dumas & Custodio, 2020).

Definir tiempos holgados genera un subdimensionamiento de las metas operativas, lo que provoca una aparente eficiencia. Esta percepción errónea, compartida entre planificadores y operarios, oculta la verdadera capacidad productiva del sistema. Desde

la perspectiva Lean, esta práctica constituye un desperdicio encubierto que frena la mejora continua y la optimización de los procesos (Ohno, 1988; Womack & Jones, 2003). La Ilustración 1 presenta evidencia empírica que respalda esta necesidad, mostrando la evolución mensual de dos indicadores clave: la precisión relativa de la planificación del tiempo de carga y el rendimiento operativo medido en sacos por hora-hombre durante el primer semestre de 2025. La literatura señala que una mayor precisión en la planificación se asocia con un uso más eficiente del tiempo y los recursos humanos, impactando positivamente en la productividad operativa (Singh, 2021; Zhou & Wang, 2019).

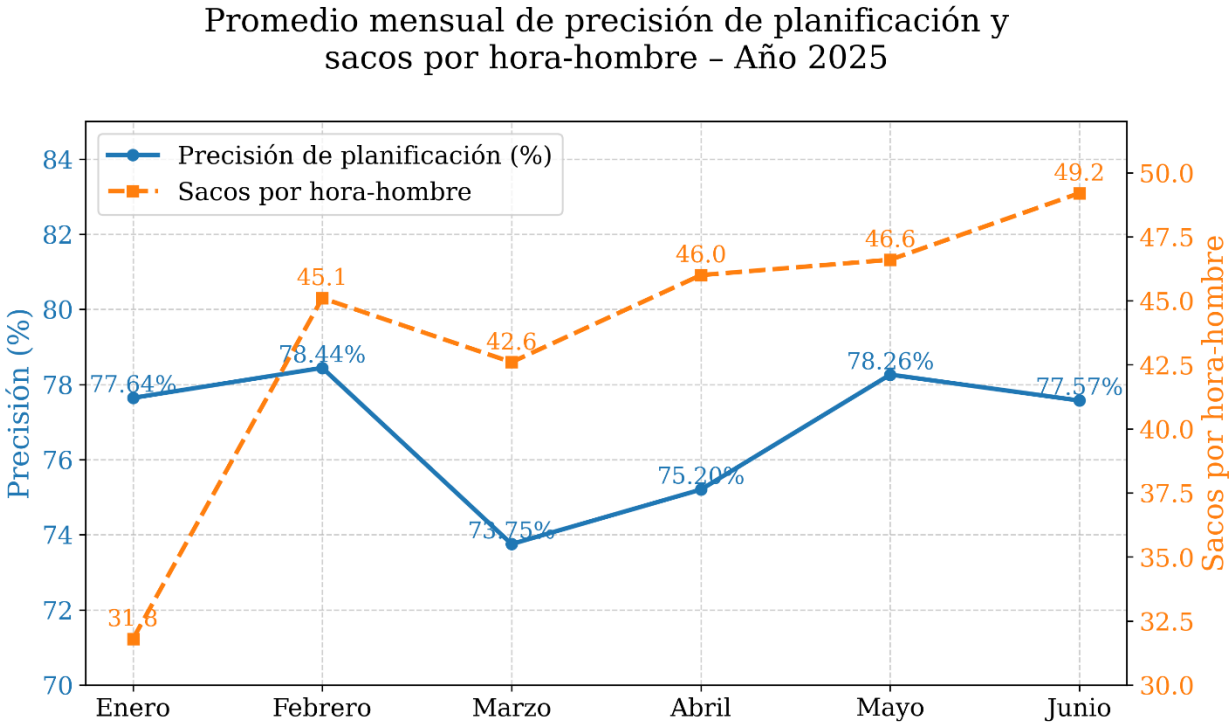

Ilustración 1 Promedio mensual de precisión de planificación y sacos por hora-hombre – Año 2025

Fuente: Elaboración propia con datos operativos de la planta de distribución de productos de nutrición de cultivos del año 2025

Como se muestra en la ilustración 1, los meses con mayor precisión en la planificación —febrero con 78,4 % y mayo con 78,26 %— presentan también valores elevados de rendimiento operativo, alcanzando 45,1 y 46,6 sacos por hora-hombre, respectivamente. En contraste, marzo reflejó una disminución en la precisión de la planificación —73,75 %— y un menor rendimiento operativo —42,6 sacos por hora-hombre—. Esta relación

sugiere que una mayor precisión en la planificación se asocia con un mejor aprovechamiento del tiempo y de los recursos humanos.

Por lo tanto, la implementación de herramientas analíticas permite monitorear estos indicadores, detectar desajustes y actuar de forma proactiva. Estratégicamente, esta propuesta permite optimizar la planificación de las cargas, mejorar la asignación de recursos humanos y físicos, y facilitar la toma de decisiones basada en evidencia. En consecuencia, impacta directamente en la eficiencia del proceso, contribuyendo al cumplimiento de cronogramas y fortaleciendo una logística más ágil, confiable y centrada en las necesidades del cliente. Asimismo, fortalece la transformación digital organizacional, alineándose con las tendencias modernas en optimización de cadenas de suministro (Araujo & Etemad, 2020).

1.3. SOLUCIÓN PROPUESTA

Se propone implementar una herramienta de visualización interactiva basada en un modelo de aprendizaje automático que estime los tiempos de carga de transportes en una planta de distribución de productos de nutrición de cultivos, con el propósito de optimizar la planificación logística y mejorar la eficiencia del proceso de carga. Este modelo utilizará variables históricas y operativas disponibles en la planta, tales como tipo de producto, volumen de carga, tiempos reales de carga históricos, línea de producción, horario de operación, entre otros. El sistema predictivo será integrado en una plataforma visual dinámica desarrollada con herramientas de visualización de datos (Power BI), que permita al equipo de planificación monitorear los tiempos proyectados versus los ejecutados, identificar desviaciones y tomar decisiones informadas. Esta interfaz permitirá también realizar simulaciones, generar alertas y establecer metas de carga basadas en datos reales. Para ello, se considerarán modelos como regresión lineal múltiple, Random Forest y XGBoost, seleccionados por su buen desempeño reportado en problemas de predicción en logística.

La solución aportará valor al mejorar la precisión en la planificación, reducir la incertidumbre operativa, optimizar la utilización de recursos y fortalecer una cultura de mejora continua apoyada en evidencia cuantitativa, contribuyendo directamente al impacto esperado del proyecto.

1.4. OBJETIVOS

- Objetivo General

Implementar una herramienta de visualización interactiva basada en un modelo de aprendizaje automático que estime los tiempos de carga de transportes en una planta de distribución de productos de nutrición de cultivos, con el propósito de optimizar la planificación logística y mejorar la eficiencia del proceso de carga.

- Objetivos Específicos
    1. Analizar los datos operativos históricos de la planta para identificar las variables más relevantes que influyen en el tiempo de carga.
    2. Implementar y entrenar un modelo de aprendizaje automático capaz de predecir de manera precisa el tiempo de carga a partir de las variables identificadas.
    3. Desarrollar una herramienta de visualización interactiva que facilite la toma de decisiones y contribuya a la optimización de los procesos logísticos en el despacho de productos.
    4. Evaluar el desempeño del modelo predictivo utilizando métricas como MAE, RMSE y R^2, comparándolo con la planificación manual actual.

1.5. METODOLOGÍA

El desarrollo del proyecto se estructura conforme a la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), ampliamente validada en proyectos de ciencia de datos por su enfoque iterativo y centrado en el negocio. Esta metodología ha sido aplicada con éxito en múltiples dominios industriales, lo que respalda su idoneidad para abordar proyectos de optimización logística basados en datos.

Esta metodología se estructura en seis fases principales:

1. Comprensión del negocio: Se analizarán los procesos operativos actuales de carga, sus limitaciones, la lógica de planificación empírica y los indicadores clave, como sacos por hora-hombre. Esta etapa permitirá definir con claridad los objetivos del proyecto y los requisitos funcionales del modelo predictivo.
2. Comprensión de los datos: Se recopilarán y evaluarán los datos históricos disponibles, incluyendo variables como fechas de carga, tipo de producto,

volumen, recursos humanos asignados, tiempos reportados y resultados operativos. Se verificará la calidad, integridad y coherencia de los datos.

3. Preparación de los datos: Se realizará la limpieza y transformación de datos, así como la selección de atributos relevantes. Esta fase incluirá la generación de nuevas variables, codificación de datos categóricos y normalización de valores temporales o numéricos.
4. Modelado: Se aplicarán algoritmos de aprendizaje supervisado, como regresión lineal múltiple, Random Forest y XGBoost para estimar el tiempo de carga esperado.
5. Evaluación: Los modelos serán evaluados con métricas como RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) y R^2 (coeficiente de determinación), tal como se muestra en la Tabla 2. Se seleccionará el modelo con mejor equilibrio entre precisión y generalización.

INDICADOR	DESCRIPCIÓN	INTERPRETACIÓN
RMSE (Root Mean Squared Error)	Raíz del error cuadrático medio	Penaliza más los errores grandes; cuanto menor, mejor
MAE (Mean Absolute Error)	Promedio de los errores absolutos	Mide el error medio sin considerar el signo; cuanto menor, mejor
R^2 (Coeficiente de determinación)	Proporción de varianza explicada por el modelo	Valores cercanos a 1 indican buen ajuste

Tabla 2 Indicadores de evaluación del modelo predictivo

6. Despliegue: El modelo seleccionado será integrado en una plataforma visual interactiva (como Power BI), la cual permitirá al usuario monitorear proyecciones, comparar con tiempos reales y facilitar la toma de decisiones operativas en tiempo real.

La ilustración 2 resume gráficamente la estructura iterativa de la metodología CRISP-DM, destacando la interacción continua entre la comprensión del negocio y el análisis de datos, lo cual garantiza la alineación del modelo predictivo con las necesidades operativas de la planta.

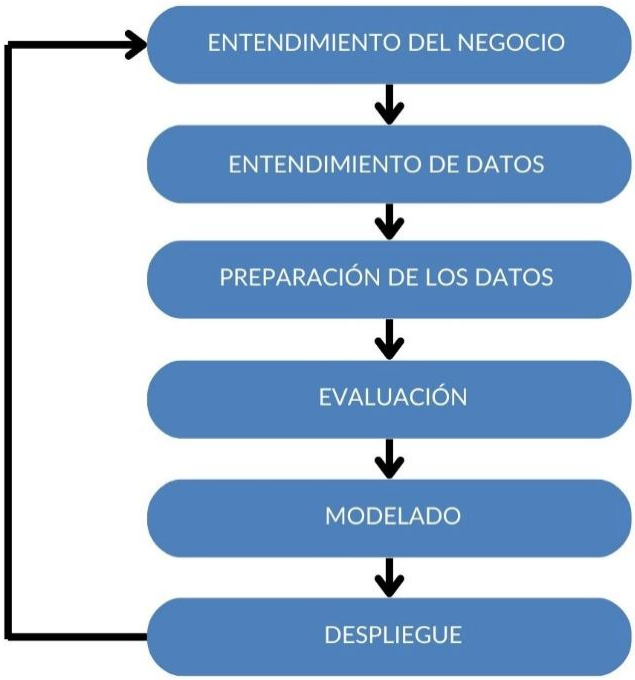

Ilustración 2 Metodología CRISP-DM

1.6. RESULTADOS ESPERADOS

Se espera que la implementación del modelo predictivo y su integración con una plataforma visual interactiva genere beneficios cuantificables en la eficiencia y planificación del proceso logístico de carga de transportes. Entre los resultados esperados se encuentra la reducción del margen de error en la estimación de tiempos de carga en, al menos, un 10 %, mediante la predicción basada en datos históricos y variables operativas clave. La literatura señala que el uso de modelos predictivos en logística permite mejorar la precisión de la planificación y reducir la incertidumbre operativa (Zhou & Wang, 2019; Singh, 2021).

De igual forma, se prevé una mejora sustancial en la precisión de la planificación, lo que permitirá una asignación más eficiente de recursos humanos y físicos. El monitoreo continuo de proyecciones frente a tiempos reales y la incorporación de alertas tempranas contribuirán a disminuir desviaciones operativas y a fortalecer una gestión más proactiva, trazable y basada en evidencia (Few, 2012; Araujo & Etemad, 2020). Finalmente, se fortalecerá la cultura de mejora continua al contar con indicadores más precisos y realistas que respalden la toma de decisiones correctivas.

Estos resultados no solo incrementarán la productividad, sino que también posicionarán a la organización hacia una gestión logística más inteligente, sostenible y alineada con las tendencias actuales de transformación digital (Araujo & Etemad, 2020; Singh, 2021; Zhou & Wang, 2019).

Ilustración 3 Resultados esperados del proyecto propuesto.

1.7. CONJUNTO DE DATOS

El presente proyecto se nutre de dos fuentes de datos principales provenientes de SAP: por un lado, la planificación de carga diaria generada por el planificador; por otro, los registros históricos reales del proceso de carga ejecutado en la planta para generar el reporte de OTIF, indicador que evalúa el cumplimiento del ciclo de carga de un transporte en un tiempo determinado, conformado por el proceso de inscripción del transporte, proceso de carga y proceso de despacho.

Los documentos fuente incluyen los archivos *"OTIF - PLANTA GYE 2024-2025.xlsx"* y, *"PLANIFICACIÓN 2024-2025.xlsx"*, los cuales contienen información operativa consolidada sobre pedidos programados desde el 1 de julio del 2024 hasta el 11 de junio del 2025, con un total de 4065 registros de cargas de transporte, donde se incluye información como tiempos de carga planificados, fecha y hora de despacho, producto a cargar, presentación de sacos (50Kg, 25 Kg, Big Bags de 800 Kg), línea de producción, si el transporte tuvo que cargar stock (es decir que cargó producto ya ensacado) y tiempos reales de carga. Esta información es generada y almacenada por el equipo de producción como parte de su rutina administrativa.

A partir de los archivos fuente se ha construido un dataset consolidado para el entrenamiento del modelo de aprendizaje automático. Este dataset está compuesto exclusivamente por variables relevantes para la predicción del tiempo de carga, extraídas de la información previamente descrita. Se han incluido variables numéricas y categóricas relacionadas con el producto, la presentación, el stock, la línea de producción y la duración real. Para el proceso de modelado se propone dividir los datos en un 80 % para el entrenamiento y un 20 % para las pruebas, garantizando así una evaluación objetiva del desempeño del modelo sobre información no utilizada durante el aprendizaje.

Durante la etapa de comprensión y preparación de los datos, se verificará la calidad del dataset, revisando valores nulos, inconsistencias, duplicados y posibles sesgos. También se aplicarán filtros para excluir registros atípicos o incompletos que puedan afectar la robustez del modelo. Se priorizará el uso de datos internos generados por la operación real, garantizando la representatividad y aplicabilidad directa de los resultados al contexto de la planta.

Año	Mes	Semana	Nº de transporte	Inicio planif.carga	Hora planif.inscrip.	Hora actual registro	Hr.prev.inicio carga	Hora act.inic.carga	Hora prev.fin carga	Hora act.fin carga	Hora plan.DespExped.	Hora act.desp.exp.	Línea producción	Peso Teórico	Peso Real
2024	11	48	80008161	11/26/2024	6:10:00 AM	4:16:47 PM	7:10:00 AM	7:10:00 AM	7:40:00 AM	7:35:00 AM	8:40:00 AM	7:59:00 AM	2	13,700	13,740
2024	1	3	80003281	1/19/2024	6:45:00 AM	6:36:17 AM	7:45:00 AM	7:00:00 AM	8:40:00 AM	7:45:00 AM	9:40:00 AM	8:00:00 AM	1	25,000	25,090
2024	1	1	80003371	1/2/2024	6:00:00 PM	11:34:50 AM	7:00:00 PM	12:17:00 PM	7:50:00 PM	1:10:00 PM	8:50:00 PM	1:16:00 PM	1	25,000	25,080
2024	1	2	80003402	1/8/2024	6:20:00 AM	6:31:50 AM	7:20:00 AM	7:20:00 AM	8:50:00 AM	9:33:00 AM	9:50:00 AM	9:49:00 AM	2	33,400	33,510
2024	1	1	80003419	1/3/2024	9:15:00 AM	8:45:21 AM	10:15:00 AM	10:10:00 AM	11:25:00 AM	11:13:00 AM	12:25:00 PM	11:25:00 AM	2	25,000	25,080
2024	1	1	80003442	1/2/2024	6:00:00 PM	4:13:06 PM	7:00:00 PM	5:00:00 PM	7:45:00 PM	5:25:00 PM	8:45:00 PM	5:37:00 PM	1	15,000	15,050
2024	1	1	80003443	1/2/2024	7:20:00 AM	7:37:52 AM	8:20:00 AM	8:20:00 AM	8:50:00 AM	10:16:00 AM	9:50:00 AM	11:48:00 AM	1	15,000	15,050
2024	1	1	80003444	1/4/2024	1:00:00 PM	8:47:19 AM	2:00:00 PM	9:35:00 AM	2:30:00 PM	9:48:00 AM	3:30:00 PM	10:06:00 AM	2	4,625	4,640

Ilustración 4 Registros de tiempos de documento OTIF - PLANTA GYE 2024-2025

Como muestra la ilustración 4, a partir de las variables registradas en el documento OTIF - PLANTA GYE 2024-2025.xlsx como Año, Mes y Semana (identificadores temporales del registro), N° de transporte (código único del vehículo o unidad que realizó la carga), Inicio planif. carga y Hora prev. inicio carga (fecha y hora previstas de inicio de la carga), Hora act. inic. carga (fecha y hora reales de inicio), Hora prev. fin carga y Hora act. fin carga (fecha y hora previstas y reales de finalización de la carga), además en este tiempo pudo haber cargado 1 o más productos, Hora plan. Desp/Exped y Hora act. desp./exp. (hora prevista y real de despacho o expedición), Línea producción (identificador de la

línea donde se realizó la carga), Peso teórico y peso real (pesos calculados y registrados para el transporte), se calcularán nuevas métricas en minutos, tales como DuracionPlanificadaMin (diferencia entre fin e inicio planificados), DuracionRealMin (diferencia entre fin e inicio reales) y delta_min (diferencia entre la duración real y la planificada), las cuales serán la base para el análisis de tiempos en el indicador OTIF.

Semana	Fecha de carga	Nombre	Cod. Material	Producto	Presentacion	Cantidad	Peso Teorico	Transporte	Línea	Fecha Inicio De Carga	Hora De Carg	Tiempo De Carg	Hora Fin de carg	ORDEN DE FABRICACION	TIPO DE PRODUCTO
2	06.01.2025	AGROAZUCAR ECUADOR S.	1E+09	MEZCLA 101/25.2N-0.0P2O5-37	K26	600.00	30,000.00	80008772	1	06.01.2025	7:10	60	8:10	200012722	Mezcla Simple
2	06.01.2025	AGROAZUCAR ECUADOR S.	1E+09	MEZCLA 101/25.2N-0.0P2O5-37	K26	600.00	30,000.00	80008773	1	06.01.2025	8:13	60	9:13	200012723	Mezcla Simple
2	06.01.2025	CEDEÑO CEDEÑO NOEL FRA	1E+09	MFFG MAIZ ESPECIAL GRA	K26	800.00	40,000.00	80008634	1	06.01.2025	9:16	85	10:41	200012687	Mezcla Compleja
2	06.01.2025	PADRON IGLESIAS PEDRO J	1E+09	30.0-0-16.0	K26	200.00	10,000.00	80008736	1	06.01.2025	10:44	25	11:09	200012787	Sin orden
2	06.01.2025	SANCHEZ TORRES ALFONS	1E+09	NITRO XTEND XP+K+S PRI 34-	K26	300.00	15,000.00	80008730	1	06.01.2025	11:12	75	12:27	200012738	Mezcla Simple
2	06.01.2025	SANCHEZ TORRES ALFONS	1E+09	NITRO XTEND XP+K+S PRI 34-	K26	500.00	25,000.00	80008730	1	06.01.2025				200012738	Mezcla Simple
2	06.01.2025	MOLINEROS GONZALEZ REI	2E+09	COMBO EXTRA PRODUCCION	U	30.00	0.00	80008697	1	06.01.2025	13:27	35	14:02	-	Sin orden
2	06.01.2025	MOLINEROS GONZALEZ REI	1E+09	NITRO XTEND XP+S GRA 40N+	K26	300.00	15,000.00	80008697	1	06.01.2025				200012739	Nitro Xtend XP
2	06.01.2025	MOLINEROS GONZALEZ REI	1E+09	NITRO XTEND FOL37.8N	L	30.00	42.30	80008697	1	06.01.2025				STOCK	Sin orden
2	06.01.2025	VACA CALERO WILLIAMS YA	1E+09	MF-G MAIZ ESPECIAL GRA	K26	60.00	3,000.00	80008728	1	06.01.2025	14:05	70	15:15	200012786	Mezcla Compleja
2	06.01.2025	SOCIEDAD CIVIL Y MERCAN	2E+09	COMBO EXTRA NITROGENO	U	20.00	0.00	80008728	1	06.01.2025				-	Sin orden
2	06.01.2025	SOCIEDAD CIVIL Y MERCAN	1E+09	NITRO XTEND XP+S PRI 40N+5	K26	200.00	10,000.00	80008728	1	06.01.2025				200012780	Nitro Xtend XP

Ilustración 5 Planificación diaria de los transportes a cargar

Como se observa en la Ilustración 5, el conjunto de datos está conformado por diversas variables relevantes para el análisis. Entre ellas se incluyen: Sem (semana del año en que se registró la operación), Fecha de carga (día en que se realizó la carga del producto), Nombre (razón social del cliente), Código (código del material en SAP), Producto (descripción del producto cargado), Presentación (tipo de empaque o capacidad), Cantidad (número de unidades cargadas), Peso teórico (peso calculado según presentación y cantidad), Transporte (identificador del vehículo asignado), Línea (número de la línea de producción utilizada), Orden de fabricación (código de la orden asociada) y Tipo de producto, variable categórica que clasifica las cargas en monoproducto (producto de un solo componente), mezcla simple (hasta tres componentes) y mezcla compleja (hasta ocho componentes). Esta última variable resulta fundamental, pues se utilizará más adelante en el modelo de predicción.

Cada fila de la base de datos corresponde a un producto cargado en un transporte específico, incorporando información operativa como la fecha de carga, pedido de venta, transporte, línea de producción, hora de inicio de carga, tiempo de carga y hora de finalización de carga. Estos atributos permiten identificar la unidad de transporte, la temporalidad de la operación y las características del producto. Cada línea del conjunto

de datos representa un producto específico asignado a un transporte, que puede corresponder tanto a una orden de fabricación como a una orden de stock. A partir de esta información es posible calcular indicadores de desempeño y analizar la relación entre los tiempos de carga y factores como la presentación, el peso, la línea y el turno de producción.

Dado que el conjunto de datos proviene de información operativa real de la empresa, se garantiza la confidencialidad y el uso exclusivo con fines académicos. Los datos han sido anonimizados y no incluyen información sensible de clientes ni detalles comerciales específicos, asegurando un manejo responsable y ético de la información.

CAPÍTULO 2

2. ESTADO DEL ARTE

El presente capítulo expone los fundamentos conceptuales, antecedentes académicos y aplicaciones relevantes relacionadas con la estimación de tiempos logísticos mediante técnicas de aprendizaje automático. Se revisan investigaciones previas que han abordado problemáticas similares, así como las metodologías y herramientas utilizadas, con el objetivo de sustentar teóricamente la propuesta planteada en este proyecto.

2.1. LOGÍSTICA Y LEAN MANUFACTURING

Según Rushton et al. (2022), la planificación logística constituye un elemento clave para el desempeño de los sistemas de distribución, ya que influye directamente en la eficiencia operativa, el uso de recursos y el cumplimiento de los tiempos establecidos. Una planificación inadecuada puede generar retrasos, incremento de costos y pérdida de competitividad en entornos industriales (Christopher, 2016; Ballou, 2004). En este contexto, la logística se consolida como un factor estratégico para la sostenibilidad y crecimiento de las organizaciones (OECD, 2020).

En entornos de distribución, la eficiencia logística está estrechamente relacionada con la correcta planificación de los procesos operativos, entre ellos el proceso de carga de transportes (Rushton et al., 2022; Ballou, 2004). La literatura señala que una planificación deficiente de los tiempos de operación genera variabilidad, cuellos de botella y un uso ineficiente de los recursos disponibles, afectando directamente el desempeño global del sistema logístico (Singh, 2021).

El enfoque Lean Manufacturing surge como una estrategia orientada a la eliminación sistemática de desperdicios y a la mejora continua de los procesos productivos (Liker, 2004). De acuerdo con Ohno (1988), los desperdicios incluyen actividades que no agregan valor al cliente, como tiempos de espera innecesarios, sobreproducción y asignaciones ineficientes de recursos. Womack y Jones (2003) destacan que la

identificación y reducción de estos desperdicios permite incrementar la eficiencia operativa y mejorar el desempeño de los procesos.

Desde la perspectiva Lean, la variabilidad no controlada en los tiempos de carga puede considerarse una fuente de desperdicio, ya que dificulta la estandarización del proceso y limita la mejora continua (Liker, 2004). En particular, la asignación de tiempos de carga basados en criterios empíricos introduce holguras artificiales que ocultan ineficiencias reales y distorsionan los indicadores de productividad, como el rendimiento medido en sacos por hora-hombre (Ohno, 1988; Womack & Jones, 2003).

En este contexto, la integración de principios Lean con enfoques basados en datos permite fortalecer la planificación logística al proporcionar estimaciones más precisas y objetivas de los tiempos operativos (Bertsimas & Kallus, 2020). Estudios recientes indican que el uso de información histórica y modelos analíticos contribuye a reducir la variabilidad, mejorar la utilización de los recursos y apoyar la toma de decisiones orientadas a la mejora continua en sistemas logísticos complejos (Araujo & Etemad, 2020; Singh, 2021).

2.2. ANTECEDENTES DEL USO DE MODELOS PREDICTIVOS EN LOGÍSTICA

Durante la última década, la adopción de modelos predictivos en logística ha crecido significativamente debido a su capacidad para anticipar eventos, optimizar recursos y reducir ineficiencias operativas (Rushton et al., 2022). A nivel internacional, Zhou y Wang (2019) emplearon Random Forest para predecir demoras en centros de distribución en China, logrando reducir tiempos de espera en un 18 %. Por su parte, Singh (2021) desarrolló modelos basados en XGBoost y KNN para estimar tiempos de llegada de camiones en India, incrementando la precisión en un 22 % frente a estimaciones tradicionales (Chen & Guestrin, 2016).

En América Latina, Araujo y Etemad (2020) integraron modelos predictivos con herramientas de visualización para anticipar retrasos en cadenas agroindustriales, alcanzando reducciones del 35 % en el error de predicción. De manera similar, Rodríguez et al. (2022) implementaron un sistema basado en árboles de decisión para estimar tiempos de carga en plantas mexicanas de fertilizantes, mejorando la eficiencia de la

planificación operativa (Breiman, 2001). Estudios regionales adicionales destacan resultados consistentes en la aplicación de analítica predictiva para procesos logísticos en industrias agroalimentarias y de insumos agrícolas (BID, 2020; CEPAL, 2019).

La literatura especializada señala que estos modelos suelen incorporar variables operativas como volumen de carga, número de operarios, tipo de producto, condiciones de despacho y características del transporte, permitiendo capturar patrones complejos que no son evidentes mediante métodos empíricos (Hastie et al., 2009). La inclusión de estas variables contribuye a generar estimaciones más precisas y consistentes, especialmente en procesos con alta variabilidad operativa, como los tiempos de carga en plantas de distribución (Singh, 2021; Zhou & Wang, 2019).

En el contexto ecuatoriano, informes institucionales han evidenciado que los procesos logísticos en sectores agroindustriales y de distribución presentan limitaciones asociadas a la planificación empírica de tiempos operativos y a la falta de herramientas analíticas para la toma de decisiones (CEPAL, 2020). Estudios realizados en el sector productivo nacional destacan que la mejora en la estimación de tiempos de operación constituye un factor clave para incrementar la eficiencia y competitividad de las plantas de distribución (Ministerio de Producción, Comercio Exterior, Inversiones y Pesca, 2021; INEC, 2022).

Asimismo, análisis desarrollados en el marco de proyectos de modernización logística en Ecuador señalan que la incorporación de enfoques basados en datos permite optimizar la planificación operativa y reducir la variabilidad en procesos de carga y despacho, especialmente en industrias vinculadas a la producción agrícola y agroindustrial (BID, 2020; CEPAL, 2020; FAO, 2021).

En términos operativos, diversos estudios coinciden en que la aplicación de modelos predictivos no solo mejora la precisión de los tiempos estimados, sino que también fortalece indicadores de desempeño como productividad, cumplimiento de programación y eficiencia del uso del recurso humano, lo que resulta fundamental para una gestión logística más confiable y orientada a la mejora continua (BID, 2020; Araujo & Etemad, 2020).

Estas investigaciones muestran cómo los modelos predictivos permiten reducir la variabilidad operativa, mejorar la precisión de la planificación y disminuir la dependencia de estimaciones empíricas (Makridakis et al., 2018). En el presente estudio, esta evidencia respalda la pertinencia de emplear técnicas de aprendizaje automático para

predecir tiempos de carga y optimizar el proceso logístico de distribución de productos de nutrición de cultivos.

2.3. INDICADORES DE EFICIENCIA LOGÍSTICA

Los indicadores logísticos permiten evaluar el rendimiento de una operación dentro de la cadena de suministro. Entre los más relevantes destacan el OTIF (On Time In Full), el nivel de servicio, el tiempo de ciclo, la productividad operativa y la exactitud de inventarios (Gattorna, 2019; Rushton et al., 2022). Estos indicadores se resumen en la Tabla 3, donde se presentan sus definiciones y fórmulas principales.

El indicador OTIF mide el cumplimiento de entregas completas y puntuales, por lo que cualquier desviación en el tiempo de carga impacta directamente su valor y, en consecuencia, la percepción de servicio al cliente. De igual modo, el indicador sacos por hora-hombre depende de contar con un tiempo de carga realista y estandarizado; estimaciones empíricas o imprecisas pueden distorsionar la medición del desempeño operativo, tal como se detalla en la Tabla 3 (Christopher, 2016).

La integración de un modelo predictivo contribuye a mejorar estos indicadores al anticipar de manera más precisa la duración de las actividades logísticas, reducir la incertidumbre y fortalecer la programación operativa (Hyndman & Athanasopoulos, 2021). En este sentido, la predicción precisa del tiempo de carga se convierte en un componente crítico para optimizar la eficiencia global del sistema.

Indicador	Definición	Fórmula
OTIF (On Time In Full)	Mide el porcentaje de entregas realizadas a tiempo y completas según lo solicitado por el cliente.	OTIF = (Entregas a tiempo y completas / Entregas totales) × 100
Nivel de servicio	Indica la capacidad de la operación para cumplir con los requerimientos del cliente sin presentar faltantes.	Nivel de servicio = (Pedidos atendidos sin faltantes / Pedidos totales) × 100
Tiempo de ciclo logístico	Tiempo total requerido para completar un proceso, desde el inicio hasta su finalización, incluyendo carga, transporte y descarga.	Tiempo de ciclo = Tiempo fin del proceso – Tiempo inicio del proceso
Productividad operativa (Sacos por hora-hombre)	Mide el rendimiento del personal en la operación de carga, calculando cuántos sacos se cargan por hora-persona.	Sacos por hora-hombre = Total de sacos cargados / (Horas trabajadas × Número de personas)
Exactitud de inventario	Evalúa la coincidencia entre las existencias registradas en el sistema y las existencias físicas.	Exactitud = (Inventario correcto / Inventario total verificado) × 100

Tabla 3 Principales indicadores logísticos

2.4. CIENCIA DE DATOS Y MACHINE LEARNING

La ciencia de datos combina técnicas estadísticas, programación y análisis de información para extraer conocimiento útil de grandes volúmenes de datos (Provost & Fawcett, 2013). Dentro de este campo, el machine learning ofrece herramientas capaces de aprender patrones complejos a partir de datos históricos para realizar predicciones o clasificaciones (Shalev-Shwartz & Ben-David, 2014). La relación entre estos conceptos y su aplicación en la predicción del tiempo de carga se representa de manera general en la ilustración 6 (Hyndman & Athanasopoulos, 2021).

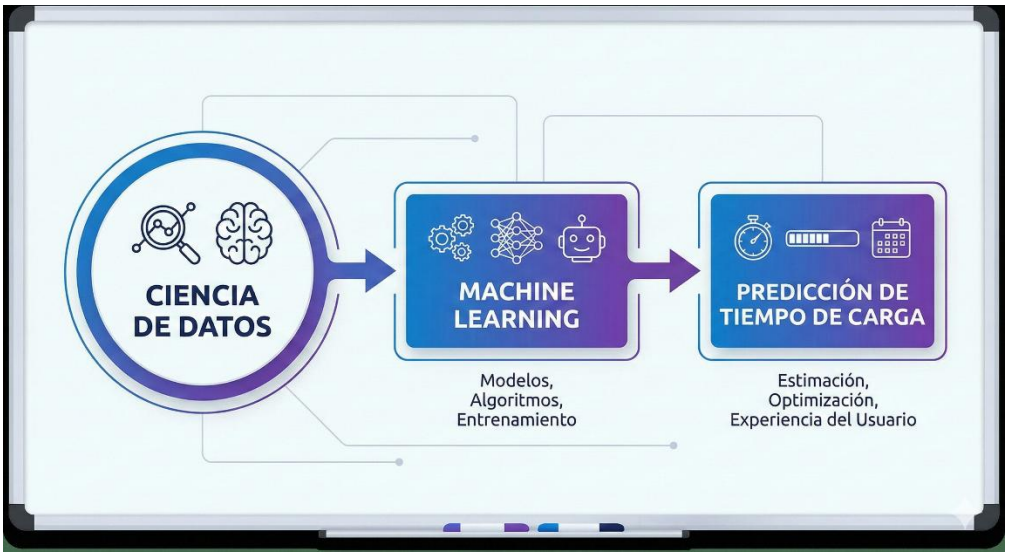

Ilustración 6 Arquitectura Lógica del Modelo de Predicción

Los modelos supervisados son especialmente adecuados para problemas de regresión, como la predicción del tiempo de carga, dado que pueden capturar relaciones no lineales entre múltiples variables operativas (Goodfellow et al., 2016). Como se resume en la Tabla 4, estos modelos presentan distintas capacidades, ventajas y limitaciones que los hacen aplicables en diferentes escenarios logísticos (Hastie et al., 2009). Estas técnicas superan a los métodos tradicionales cuando existen interacciones entre variables, datos con ruido o patrones operativos no evidentes para el analista humano (Shalev-Shwartz & Ben-David, 2014).

Para el presente proyecto, el machine learning permite identificar los factores que influyen en la duración del proceso de carga y estimar el tiempo de manera más precisa, apoyando decisiones operativas basadas en evidencia.

Modelo	Tipo	Ventajas	Desventajas	Cuándo usarlo
Regresión Lineal	Lineal	Fácil de interpretar, rápido, base de comparación.	No captura relaciones no lineales, sensible a outliers.	Cuando las relaciones entre variables son simples y lineales.
Árboles de Decisión	No lineal	Interpretación visual, puede manejar interacciones y no linealidades.	Tiende al sobreajuste si no se regula.	Cuando se necesita explicar la lógica del modelo.
Random Forest	Ensamble	Alta precisión, robusto al ruido y outliers, maneja no linealidad.	Difícil de interpretar, más demandante computacionalmente.	Cuando se busca buen desempeño general sin necesidad de full interpretabilidad.
Gradient Boosting (XGBoost, LightGBM)	Ensamble	Excelente precisión, captura patrones complejos, eficiente en datos tabulares.	Requiere tuning, menor interpretabilidad.	Cuando se busca el mejor rendimiento predictivo.
Red Neuronal Feedforward	No lineal	Excelente para relaciones complejas y datos con múltiples variables.	Necesita más datos, menor interpretabilidad y mayor tiempo de entrenamiento.	Cuando existen patrones difíciles de capturar con modelos tradicionales.
LSTM (Redes recurrentes)	No lineal secuencial	Captura dependencias temporales y dinámicas en secuencias.	Requiere más recursos, configuración más compleja.	Cuando el tiempo de carga depende de secuencias o tendencias temporales.

Tabla 4 Modelos de Regresión

2.5. APLICACIÓN DE LA METODOLOGÍA CRISP-DM EN PROYECTOS DE CIENCIA DE DATOS

CRISP-DM (Cross-Industry Standard Process for Data Mining) constituye una de las metodologías más empleadas en ciencia de datos debido a su estructura clara y adaptable a múltiples industrias (Chapman et al., 2000). Comprende seis fases:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Despliegue

Landín y Reina (2021) aplicaron CRISP-DM en una planta logística y lograron mejorar la precisión de la programación operativa en un 12 %. Plotnikova et al. (2022) identificaron brechas relacionadas con privacidad, trazabilidad y adecuación metodológica al implementar CRISP-DM en una institución financiera europea. Asimismo, Brzozowska et

al. (2023) demostraron que la fase más demandante de la metodología fue la preparación de datos, que representó el 45 % del esfuerzo total en un proyecto de predicción de tiempos de ensamblaje tal como se muestra en la ilustración 7.

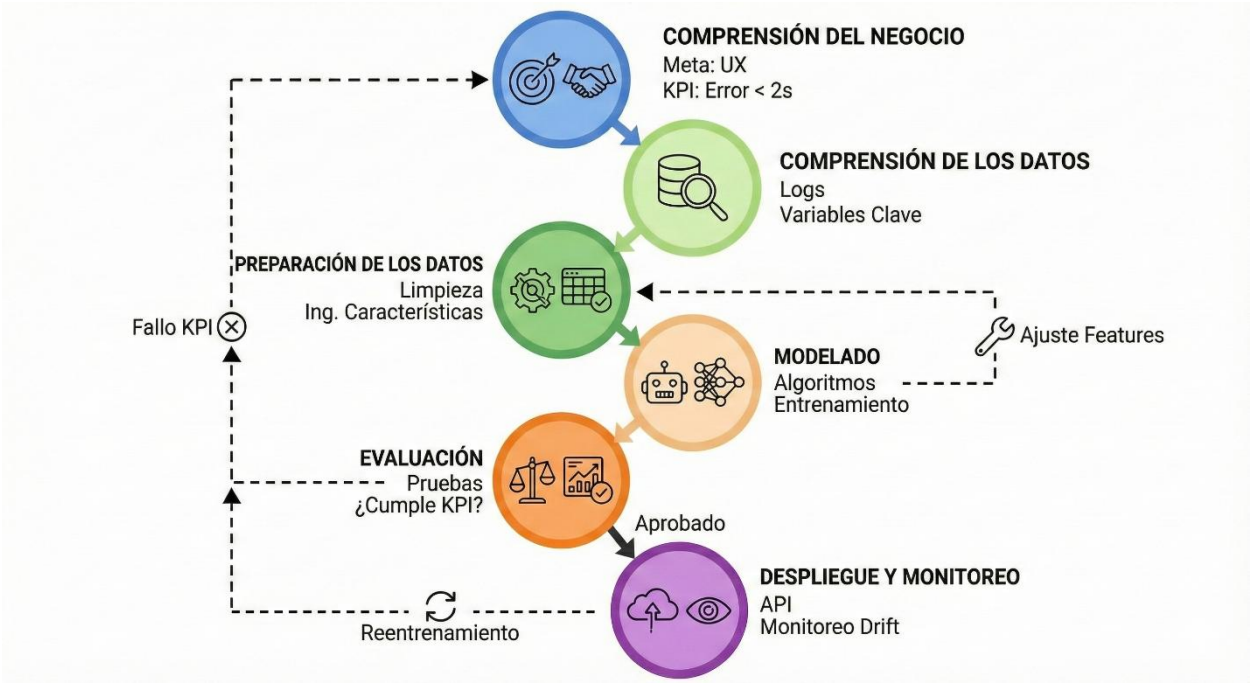

Ilustración 7 Diagrama de flujo CRISP-DM

En este estudio, CRISP-DM proporciona un marco estructurado para desarrollar el modelo predictivo del tiempo de carga. La comprensión del negocio identifica los factores operativos relevantes; la comprensión y preparación de datos permite depurar y seleccionar variables; el modelado facilita la comparación de algoritmos como regresión lineal, Random Forest y XGBoost; y finalmente, el despliegue en Power BI permite la integración del modelo con la planificación operativa diaria (Chapman et al., 2000; Landín & Reina, 2021).

2.6. MODELOS DE MACHINE LEARNING UTILIZADOS

● REGRESIÓN LINEAL MÚLTIPLE

Modelo estadístico clásico que evalúa la relación entre variables independientes y una variable dependiente. Se utiliza como línea de base para comparar el desempeño de modelos más avanzados (Montgomery et al., 2015).

● RANDOM FOREST

Algoritmo de ensamble basado en múltiples árboles de decisión que mejora la generalización del modelo y maneja adecuadamente relaciones no lineales (Breiman, 2001).

● XGBoost

Algoritmo de *gradient boosting* que destaca por su eficiencia computacional y su alto rendimiento en problemas de regresión industrial (Chen & Guestrin, 2016).

La selección de estos modelos permite comparar enfoques lineales y no lineales, identificar patrones en los datos operativos y seleccionar la técnica que brinde mayor precisión en la predicción del tiempo de carga (Hastie et al., 2009; Breiman, 2001)

Modelo	Tipo	Ventaja principal	Limitación	Referencia
Regresión Lineal	Modelo paramétrico supervisado	Interpretación sencilla y rápida; útil como línea base	No captura relaciones no lineales	Montgomery et al. (2015)
Random Forest	Ensamble de árboles (bagging)	Alta capacidad de generalización; maneja no linealidades y variables correlacionadas	Menor interpretabilidad que modelos simples	Breiman (2001)
XGBoost	Ensamble de árboles (boosting)	Excelente rendimiento en datasets tabulares; optimización interna avanzada	Mayor riesgo de sobreajuste si no se regula	Chen & Guestrin (2016)

Tabla 5 Comparación de modelos de regresión utilizados en el estudio

En el contexto del presente estudio, la variable dependiente corresponde al tiempo total de carga de los transportes (expresada en minutos), mientras que las variables independientes incluyen factores operativos como el peso real del producto cargado, la presentación (K06, K16, K26, etc.), el número de unidades por lote, el horario de carga y el día de operación. Estos atributos permiten contextualizar la aplicación de los modelos de regresión evaluados y explicitar cómo cada algoritmo busca capturar relaciones entre los factores operativos y la duración del proceso logístico (Hastie et al., 2009).

2.7. VISUALIZACIÓN DE DATOS Y POWER BI

La visualización de datos facilita la interpretación de información compleja a través de gráficos, paneles interactivos y herramientas que permiten identificar tendencias, anomalías y relaciones entre variables (Few, 2012).

Power BI es una plataforma de análisis empresarial que permite integrar modelos predictivos, automatizar la actualización de datos, crear dashboards interactivos y facilitar el acceso a información en tiempo real (Microsoft, 2023).

En este proyecto, Power BI cumple un rol fundamental al permitir que los resultados del modelo predictivo sean utilizados por el personal operativo, comparando tiempos estimados vs. reales, identificando desviaciones y apoyando la toma de decisiones diarias *(Few, 2012; Microsoft, 2023)*.

El análisis del estado del arte evidencia que el uso de modelos predictivos basados en aprendizaje automático, junto con metodologías estructuradas como CRISP-DM y herramientas de visualización como Power BI, constituye un marco sólido para mejorar la eficiencia en la planificación logística. Las investigaciones revisadas muestran mejoras significativas en precisión predictiva y optimización operativa, lo cual respalda la pertinencia del enfoque adoptado en este proyecto (Araujo & Etemad, 2020; Brzozowska et al., 2023; BID, 2020).

Con base en estos fundamentos teóricos, el siguiente capítulo presenta la metodología aplicada y el desarrollo del modelo para estimar los tiempos de carga de transportes en la planta de productos de nutrición de cultivos.

Para reforzar la comprensión visual del enfoque aplicado, se incluye la ilustración 8, donde se muestra un dashboard elaborado en Power BI. Esta representación permite observar cómo se disponen los indicadores operativos clave, tales como el tiempo estimado de carga, el tiempo real ejecutado, la diferencia entre ambos y alertas visuales que facilitan la toma de decisiones. Aunque la ilustración es de carácter conceptual, refleja la estructura general que adopta la herramienta desarrollada en este proyecto durante el proceso de despliegue de la metodología CRISP–DM (Chapman et al., 2000; Plotnikova et al., 2022).

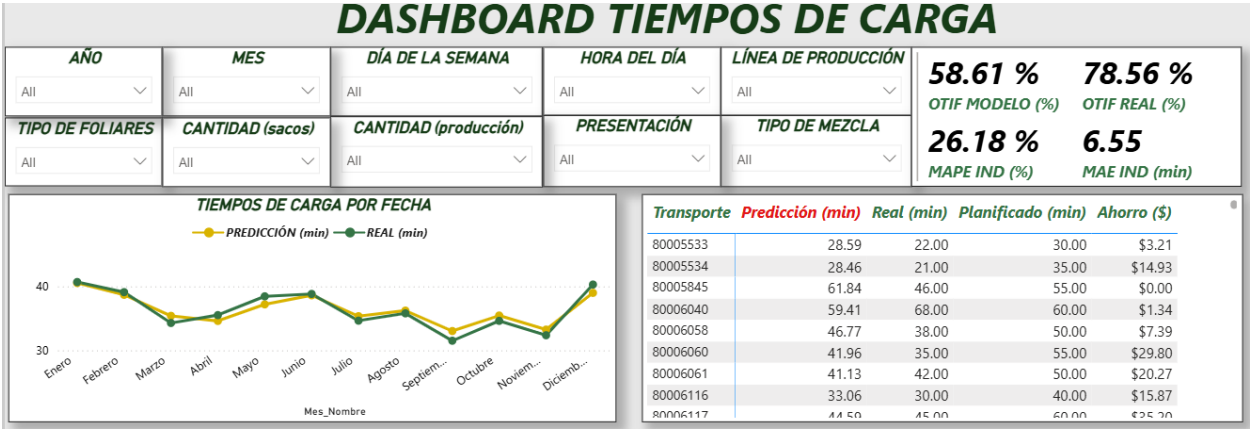

Ilustración 8 Ejemplo conceptual de un dashboard en Power BI para la estimación de tiempos de carga

CAPÍTULO 3

3. ESQUEMA GENERAL DE IMPLEMENTACIÓN

El presente capítulo describe el desarrollo de los distintos modelos propuestos para la estimación de los tiempos de carga de transporte, incluyendo el análisis de sus métricas de desempeño. Finalmente, se presenta la integración del modelo seleccionado en una herramienta digital diseñada para ser utilizada por el planificador de producción, con el objetivo de optimizar la toma de decisiones operativas.

El diseño del proyecto se estructuró en dos componentes fundamentales, que permiten abordar de manera organizada tanto la construcción del modelo como su aplicación práctica

1. Desarrollo del pipeline para el entrenamiento del modelo.
2. Desarrollo del pipeline para la implementación del modelo en una herramienta digital funcional.

Ambos componentes son esenciales para garantizar el cumplimiento de los objetivos planteados. Cada uno involucra una serie de etapas que aseguran la correcta ejecución y operatividad del sistema propuesto.

3.1. ENTENDIMIENTO DEL NEGOCIO

La gestión de los tiempos de carga constituye un punto crítico en la operación logística de la planta, ya que impacta directamente en la eficiencia de las líneas de producción y en el cumplimiento de compromisos con los clientes. Una estimación inexacta de estos tiempos puede derivar en retrasos, uso ineficiente de recursos y desajustes en la programación de transportes. Dado que la planta procesa en promedio alrededor de 600 transportes mensuales, una desviación de 15–20 minutos por transporte puede traducirse en varias decenas de horas operativas adicionales al mes, con impacto directo en costos de mano de obra y uso de infraestructura.

Por ello, el negocio requiere contar con un sistema que le permita anticipar la duración de las cargas de manera confiable, a fin de respaldar las decisiones de planificación y reducir la dependencia de estimaciones subjetivas. La solución debe alinearse con

indicadores clave como OTIF y la optimización del desempeño operativo, aportando una base cuantitativa para mejorar la programación y la asignación de recursos.

La diferencia entre el tiempo planificado y real tiene impacto directo en los costos operativos de la planta, considerando un equipo de 7 operarios por línea. Este costo se utiliza posteriormente para evaluar los beneficios del modelo predictivo.

Esta alineación entre eficiencia operativa y precisión de planificación coincide con la relación señalada por Dumas & Custodio (2020), quienes destacan que la variabilidad en los tiempos logísticos impacta directamente en indicadores como OTIF y productividad

3.2. PREPARACIÓN DE LOS DATOS

La fuente de datos correspondiente a la planificación y al indicador OTIF, descrita en la sección 1.7, fue incorporada al análisis en Python 3.10, empleando las librerías pandas, numpy y datetime. Los archivos originales fueron convertidos a dataframes, estructuras tabulares que permiten manipular la información en filas y columnas de forma eficiente.

En esta etapa se realizaron las siguientes tareas:

- Estandarización de nombres de columnas y tipos de datos: el campo Transporte se convirtió a tipo numérico entero, mientras que las fechas y horas se transformaron al formato datetime.
- Eliminación de duplicados e inconsistencias: se descartaron registros repetidos y transportes sin orden de fabricación válida, así como cuatro casos con horarios fuera del rango operativo (06:00–21:00), los cuales correspondían a errores administrativos
- Filtrado de valores atípicos evidentes en los tiempos de carga, únicamente cuando afectaban la coherencia temporal de la información.

Como resultado, el conjunto de datos de OTIF quedó conformado por 6688 filas y 14 columnas tal como se muestra en la ilustración 9.

Año	Mes	Semana	Nº de transporte	Inicio planif.carga	Línea producción	Peso Teórico	Peso Real	InicioPlanif	FinPlanif	InicioReal	FinReal	Duracion Planificada Min	Duracion Real Min
2024	11	48	80008161	2024-11-26 00:00:00	2	13700	13740	2024-11-26 07:10:00	2024-11-26 07:40:00	2024-11-26 07:10:00	2024-11-26 07:35:00	30	25
2024	1	3	80003281	2024-01-19 00:00:00	1	25000	25090	2024-01-19 07:45:00	2024-01-19 08:40:00	2024-01-19 07:00:00	2024-01-19 07:45:00	55	45
2024	1	1	80003371	2024-01-02 00:00:00	1	25000	25080	2024-01-02 19:00:00	2024-01-02 19:50:00	2024-01-02 12:17:00	2024-01-02 13:10:00	50	53
2024	1	2	80003402	2024-01-08 00:00:00	2	33400	33510	2024-01-08 07:20:00	2024-01-08 08:50:00	2024-01-08 07:20:00	2024-01-08 09:33:00	90	133
2024	1	1	80003419	2024-01-03 00:00:00	2	25000	25080	2024-01-03 10:15:00	2024-01-03 11:25:00	2024-01-03 10:10:00	2024-01-03 11:13:00	70	63
2024	1	1	80003442	2024-01-02 00:00:00	1	15000	15050	2024-01-02 19:00:00	2024-01-02 19:45:00	2024-01-02 17:00:00	2024-01-02 17:25:00	45	25
2024	1	1	80003443	2024-01-02 00:00:00	1	15000	15050	2024-01-02 08:20:00	2024-01-02 08:50:00	2024-01-02 08:20:00	2024-01-02 10:16:00	30	116
2024	1	1	80003444	2024-01-04 00:00:00	2	4625	4640	2024-01-04 14:00:00	2024-01-04 14:30:00	2024-01-04 09:35:00	2024-01-04 09:48:00	30	13
2024	1	1	80003445	2024-01-03 00:00:00	1	11800	11830	2024-01-03 18:10:00	2024-01-03 18:50:00	2024-01-03 16:45:00	2024-01-03 17:20:00	40	35
2024	1	1	80003446	2024-01-03 00:00:00	1	27500	27590	2024-01-03 16:40:00	2024-01-03 18:00:00	2024-01-03 14:05:00	2024-01-03 15:22:00	80	77

Ilustración 9 Conjunto de datos OTIF

Como resultado del proceso de preparación, el conjunto de datos de planificación está compuesto por 11.807 registros distribuidos en 8 variables, según se observa en la Ilustración 10.

Cantidad	Producto	Orden de fabricación	Transporte	Tipo de producto	Cod. Material	Presentacion	Nombre
600	ETIQ: 26,9N 0.0P 24,9K2O 0.0S 0.0Zn	200008966	80006117	Mezcla Simple	1003001077	K26	AGROAZUCAR ECUADOR S.A
600	ETIQ: 33.3N 0.0P2O5 16.6K2O	200008967	80006118	Mezcla Simple	1003001077	K26	AGROAZUCAR ECUADOR S.A
600	ETIQ: 28.1N 0.0P2O5 23.4K2O	200008968	80006119	Mezcla Simple	1003001077	K26	AGROAZUCAR ECUADOR S.A
84	MF-G ESPECIAL GRA 50kg	200008965	80006191	Mezcla Compleja	1003000131	K26	CULTIVAGRO S.A.
294	MF-G ESPECIAL GRA 50kg	200008969	80006201	Mezcla Simple	1003000131	K26	AGROTAGROW S.A.
350	MF-G 0-0-30+10MgO+8S 50 kg	200008972	80006189	Mezcla Simple	1003000465	K26	VIVANCO AGROPRODUCTORES S.C.C
150	MFFG 14,6-5-27+3MgO+2,4S+0,01B+0,03Zn 50 kg	200008971	80006189	Mezcla Simple	1003000892	K26	VIVANCO AGROPRODUCTORES S.C.C
10	PRECISAGRO BORO 1L EC	STOCK	80006189	Sin Orden	1003003194	L	HERRERA CRIOLLO MIGUEL ANGEL
566	FORMULA -MFFG BANANO ESPECIAL SUMIFRU	200008970	80006131	Mezcla Compleja	1003000709	K26	SUMIBANANAS S.A
197	FORMULA -MFFG BANANO ESPECIAL SUMIFRU	200008970	80006131	Mezcla Compleja	1003000709	K26	SUMIBANANAS S.A

Ilustración 10 Conjunto de datos de planificación

- Unión de conjunto de datos OTIF + PLANIFICACIÓN

La unión de las bases de datos se realizó mediante un INNER JOIN utilizando como clave el campo Transporte, lo que permitió consolidar en un único dataset denominado merge1 la información del transporte junto con el detalle de los productos cargados. Posteriormente, con el objetivo de evitar duplicidades y reducir la cardinalidad, los registros asociados a un mismo transporte fueron agrupados en una sola fila, preservando los valores representativos de cada variable.

Como resultado de este proceso de integración y depuración, se obtuvo un conjunto de datos con las siguientes variables:

Variable	Descripción
Transporte	Identificador o código del transporte utilizado
Año	Año en que se realizó la carga
Mes	Mes correspondiente al registro de carga
Semana	Número de semana del año
Inicio_planif.carga	Fecha y hora planificada de inicio de carga
Línea producción	Código o nombre de la línea de producción
Peso Teórico	Peso teórico calculado según la planificación
Peso Real	Peso real medido durante la carga
InicioPlanif	Fecha y hora de inicio planificada del proceso
FinPlanif	Fecha y hora de finalización planificada del proceso
InicioReal	Fecha y hora real de inicio del proceso
FinReal	Fecha y hora real de finalización del proceso
DuracionRealMin	Duración real del proceso de carga en minutos
cantidad_stock_litros	Cantidad disponible en stock en litros
cantidad_stock_L08	Cantidad de envases de foliares de 5L (productos líquidos aplicados sobre las hojas para nutrir o corregir deficiencias) en stock a cargar
cantidad_stock_L11	Cantidad de envases de foliares DE 10L (productos líquidos aplicados sobre las hojas para nutrir o corregir deficiencias) en stock a cargar
cantidad_saco_stock	Cantidad de sacos de 50 kg en stock a cargar en transporte
cantidad_produccion_k06_mezcla_complej	Producción del tipo K06 (saco de 5 kg) formulada con mezcla compleja
cantidad_produccion_k06_mezcla_simple	Producción del tipo K06 (saco de 5 kg) formulada con mezcla simple
cantidad_produccion_k06_monoproducto	Producción del tipo K06 (saco de 5 kg) formulada como monoproducto
cantidad_produccion_k16_mezcla_complej	Producción del tipo K16 (saco de 25 kg) formulada con mezcla compleja
cantidad_produccion_k16_mezcla_simple	Producción del tipo K16 (saco de 25 kg) formulada con mezcla simple
cantidad_produccion_k16_monoproducto	Producción del tipo K16 (saco de 25 kg) formulada como monoproducto
cantidad_produccion_k26_mezcla_complej	Producción del tipo K26 (saco de 50 kg) formulada con mezcla compleja
cantidad_produccion_k26_mezcla_simple	Producción del tipo K26 (saco de 50 kg) formulada con mezcla simple
cantidad_produccion_k26_monoproducto	Producción del tipo K26 (saco de 50 kg) formulada como monoproducto
cantidad_produccion_k62_mezcla_complej	Producción del tipo K62 (big bag de 800 kg) formulada con mezcla compleja
cantidad_produccion_k62_mezcla_simple	Producción del tipo K62 (big bag de 800 kg) formulada con mezcla simple
cantidad_produccion_k62_monoproducto	Producción del tipo K62 (big bag de 800 kg) formulada como monoproducto
cantidad_produccion_kg_mezcla_compleja	Producción total (big bag de 1000 kg) formulada con mezcla compleja
cantidad_produccion_kg_mezcla_simple	Producción total (big bag de 1000 kg) formulada con mezcla simple
cantidad_produccion_kg_monoproducto	Producción total (big bag de 1000 kg) formulada como monoproducto

Ilustración 11 Variables del conjunto de datos consolidados

Además, para facilitar la interpretación de las variables relacionadas con las presentaciones de los productos, se estableció una tabla de referencia de códigos que detalla el tipo de empaque, su equivalencia en peso o volumen, y su uso dentro del proceso productivo o de despacho, tal como muestra la tabla 6.

Esta clasificación permite distinguir claramente entre las presentaciones que forman parte del proceso de producción efectiva y aquellas que se consideran únicamente para control de stock o despacho.

Código	Descripción	Observación
K06	Saco de 5 kg	Producción
K16	Saco de 25 kg	Producción
K26	Saco de 50 kg	Producción
K53	Big bag de 1.200 kg	Producción
K62	Big bag de 800 kg	Producción
K73	Saco de 9,07 kg	Producción
KG	Big bag de 1.000 kg	Producción
L08	Envase de 5 litros	Solo stock
L11	Envase de 10 litros	Solo stock
L	Envase de 1 litro	Solo stock

Tabla 6 Tabla de presentación de productos

En la tabla 6 se muestra que la carga se mide en kilogramos (códigos que tienen la letra K) y puede presentarse en formato de saco tradicional o mochila; en los códigos L, la unidad es el litro.

Por ejemplo, el transporte 80006040 registró una carga conformada por:

Producto	Descripción	Tipo de formulación	Cantidad	Unidad
K26	Saco de 50 kg	Mono producto	800	sacos
K26	Saco de 50 kg	Mezcla compleja	300	sacos
L08	Envase de 5 litros	Stock	200	litros

Tabla 7 Ejemplo ilustrativo de carga de transporte 80006040

En el conjunto de datos consolidado, la información presentada en la Tabla 7 se encuentra representada en una única fila por transporte. En este formato, las cantidades correspondientes al producto K26 en sus variantes de monoproducto y mezcla compleja se registran en las columnas cantidad_produccion_k26_monoproducto y cantidad_produccion_k26_mezcla_compleja, respectivamente. De igual forma, los volúmenes asociados al producto L08 se almacenan en la columna cantidad_stock_L08.

- Análisis exploratorio guiado a modelado

Durante el análisis de los tiempos planificados y reales se identificaron 37 registros con valores atípicos en la variable DuracionRealMin, correspondientes principalmente a casos en los que el proceso operativo sufrió interrupciones o reprocesos como por ejemplo caída del servidor o no tener acceso al internet o el cliente no tiene cupo así que por tal razón se descargó el transporte, es decir se bajan los sacos cargados del transporte. Debido a que estos valores no representan la dinámica típica del proceso, fueron excluidos del análisis exploratorio para evitar que sesgaran la visualización y el cálculo de estadísticas descriptivas.

Para analizar la distribución entre la duración planificada y la real de las cargas en minutos, se elaboró un diagrama de caja y un histograma superpuesto de distribuciones de las variables DuracionPlanificadaMin y DuracionRealMin, con el objetivo de identificar posibles sesgos sistemáticos derivados de una subestimación o sobreestimación en la planificación, así como evaluar la dispersión de los tiempos registrados. A continuación, se muestran las ilustraciones para su posterior análisis.

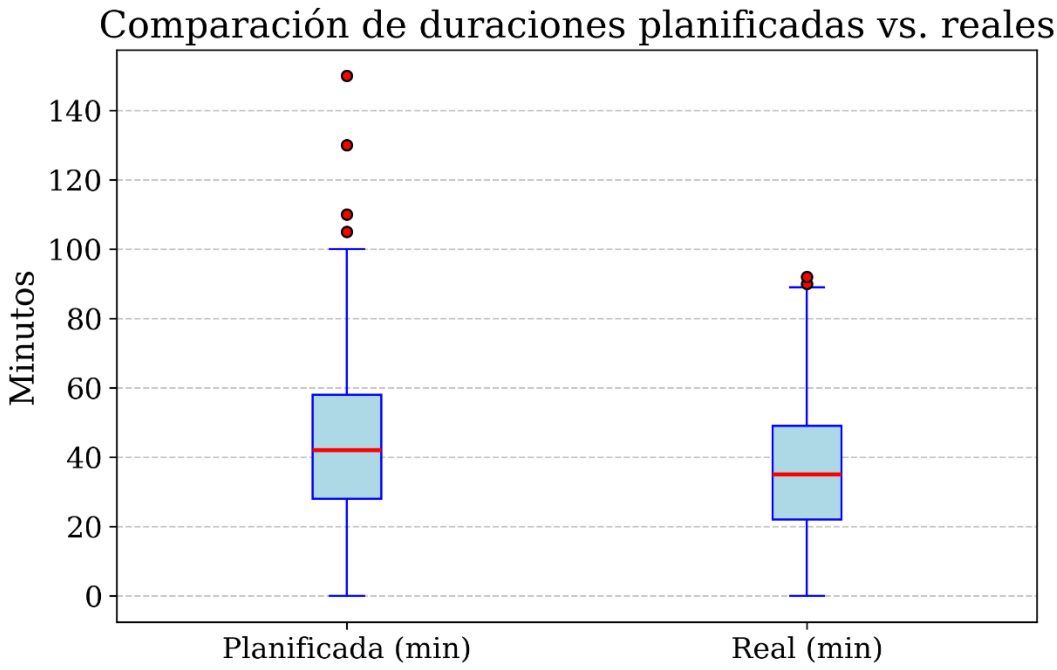

Ilustración 12 Diagrama de caja de tiempos planificados vs tiempos reales

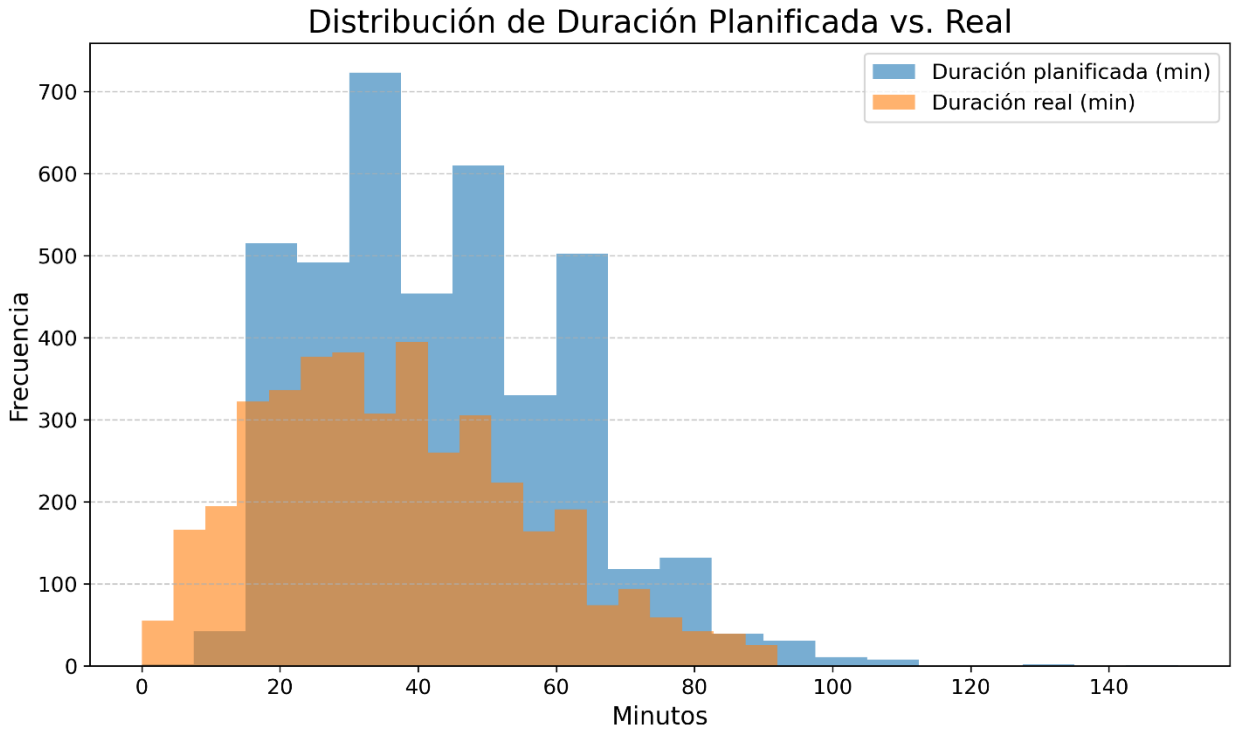

Ilustración 13 Distribución de tiempos planificados y tiempos reales de carga

El análisis comparativo de las ilustraciones 12 y 13 muestra que la distribución de los tiempos reales se concentra más cerca de los tiempos planificados y, en la mayoría de los casos, se sitúa por debajo de estos. La mediana de la duración real resulta inferior a la de la duración planificada, lo que indica que, sin considerar incidencias excepcionales, la planificación tiende a estar sobredimensionada, asignando más tiempo del realmente necesario para la carga. Además, se identificó la necesidad de tratar los valores atípicos en DuracionRealMin, dado que esta variable refleja condiciones reales sujetas a variabilidad, mientras que la DuracionPlanificadaMin se considera una referencia ideal; por tal motivo, no se considerará en el análisis exploratorio.

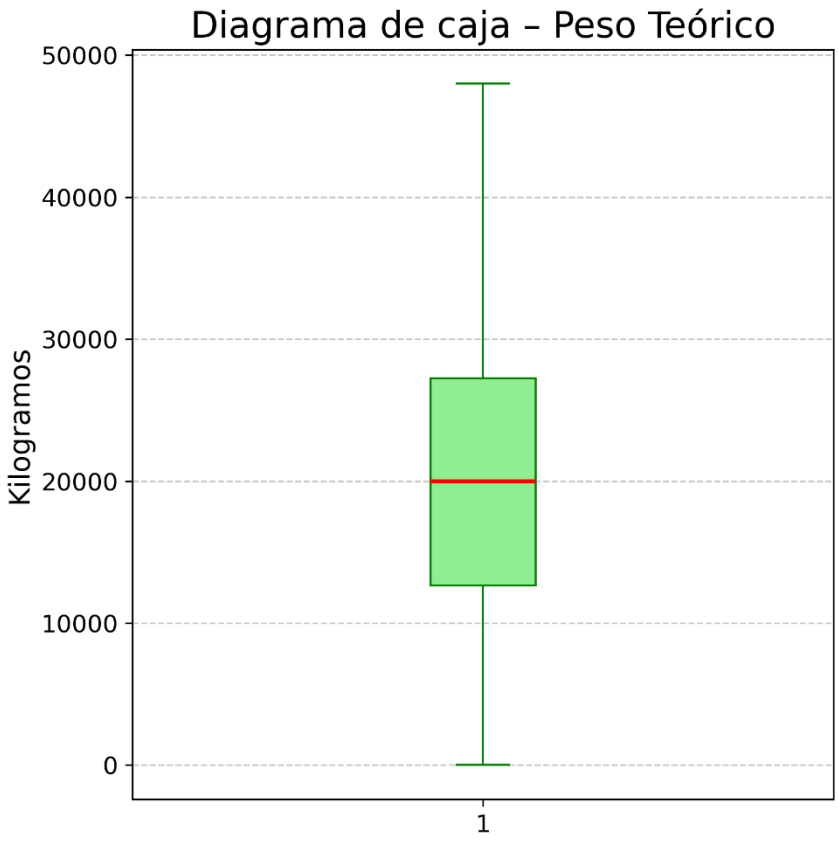

Ilustración 14 Diagrama de caja de peso teórico

El diagrama de cajas de Peso Teórico de la ilustración 14 muestra que la mayoría de los transportes se concentran en un rango relativamente estrecho de peso, con una mediana cercana a los valores máximos permitidos para la capacidad estándar de carga, lo que sugiere un uso eficiente del espacio y la capacidad de los vehículos en la mayoría de los casos. Sin embargo, se observan varios valores atípicos por debajo del rango intercuartílico inferior, correspondientes a transportes con cargas significativamente menores a la capacidad promedio.

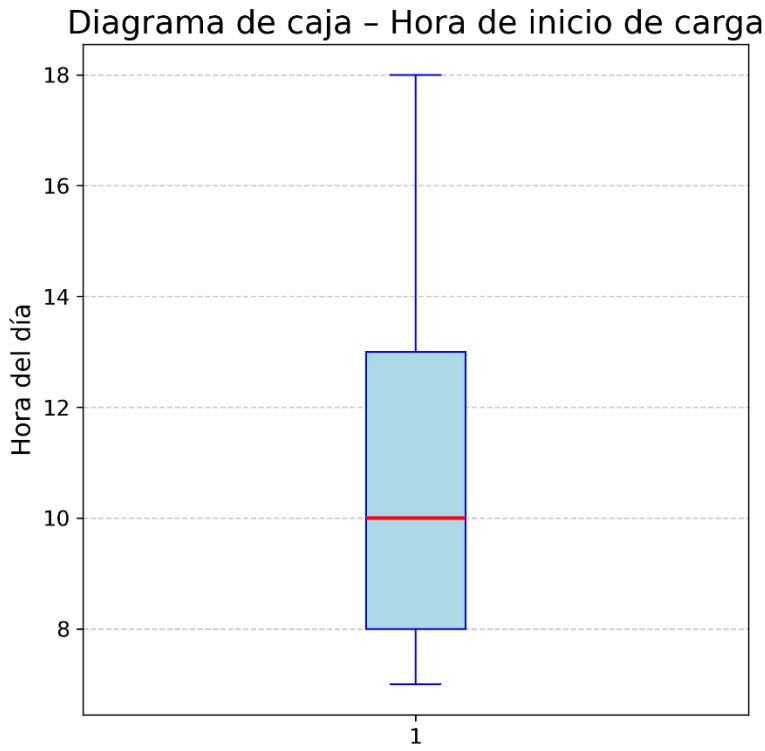

Ilustración 15 Diagrama de cajas de hora inicio de carga

La ilustración 15 sugiere que la mayoría de las cargas inician dentro de un rango horario operativo bien definido, concentrado en las horas centrales de la jornada laboral, lo que refleja una planificación estable y consistente en la programación de las operaciones. La ausencia de valores atípicos indica que, tras filtrar casos excepcionales, no existen cargas iniciadas en horarios atípicos como madrugada o noche, lo cual es coherente con un flujo de trabajo estandarizado y posiblemente alineado con los turnos de producción y disponibilidad de personal. Adicionalmente en la variable InicioReal_hora no se identificaron outliers significativos.

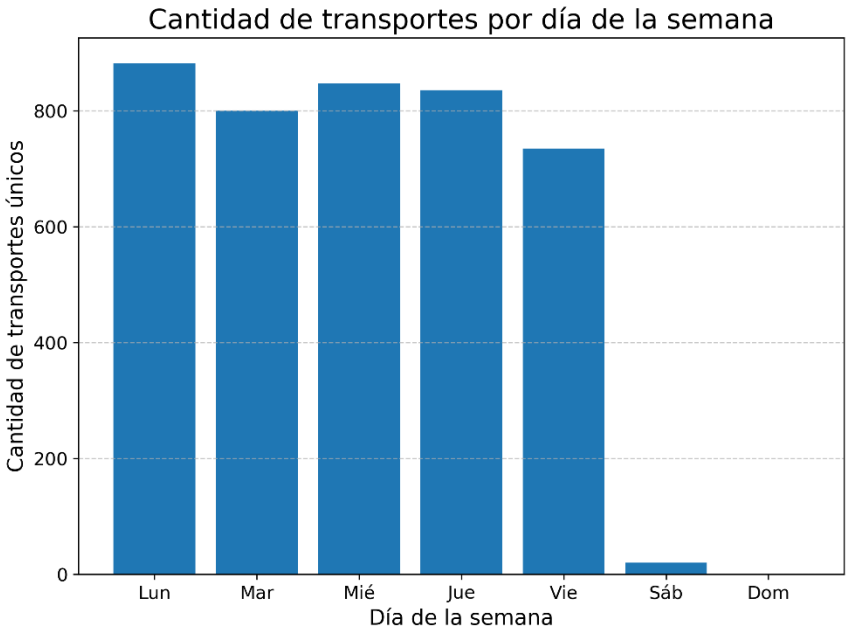

Ilustración 16 Cantidad de transportes por día de semana

La ilustración 16 muestra que la mayor cantidad de transportes se concentra de lunes a jueves, con un ligero descenso el viernes y una caída drástica el sábado, mientras que el domingo no se registran operaciones.

Esto sugiere que la operación logística está organizada principalmente en días laborales estándar, optimizando la carga y despacho en el inicio y mitad de la semana. La disminución el viernes y sábado e indica un cierre progresivo de las actividades, para dar paso a labores de mantenimiento, inventario y limpieza

En términos de planificación, este patrón permite prever picos de trabajo de lunes a jueves, donde sería clave asegurar disponibilidad plena de recursos humanos y de transporte, mientras que los días de menor actividad podrían destinarse a ajustes operativos o tareas complementarias.

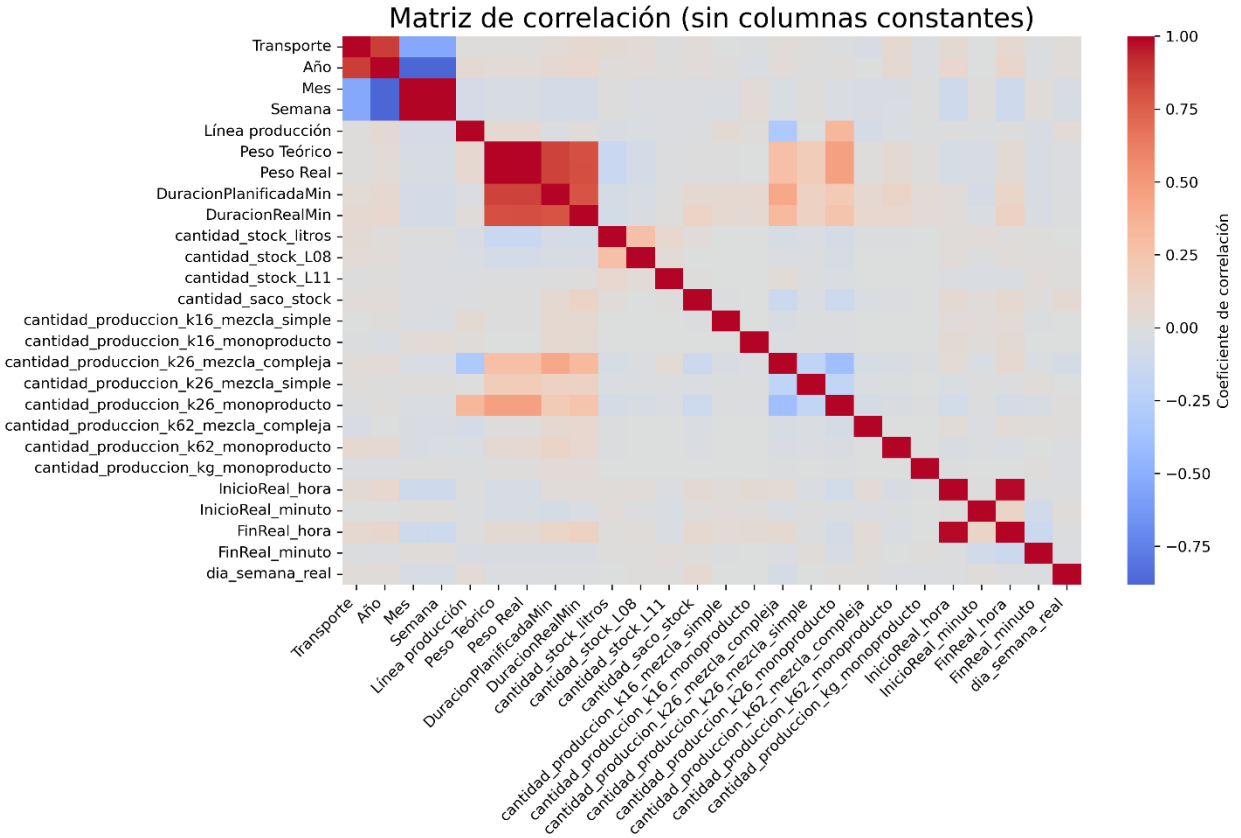

Ilustración 17 Matriz de correlación de variables

De la ilustración 17 se puede observar de la matriz de correlaciones que existen variables con multicolinealidad alta, es decir, correlaciones con magnitudes cercanas a 1

- Peso Teórico y Peso Real: Se puede observar una correlación prácticamente perfecta, lo que indica que miden casi la misma información. Mantener ambas en un modelo puede ser redundante, por eso consideraremos *PESO TEÓRICO* como característica de nuestro conjunto de datos para los modelos de predicción.
- DuracionPlanificadaMin y DuracionRealMin: aunque no parecen ser perfectamente correlacionadas, sí presentan una relación positiva considerable, lo que indica que cuando se planifica más tiempo, normalmente el tiempo real también aumenta.
    - Sin embargo, es importante mencionar que DuracionRealMin es nuestra variable a predecir y DuracionPlanificadaMin no se debe considerar debido a que esta es la variable que genera el planificador de manera heurística entonces descartamos estas variables

- Variables de producción por tipo de producto y presentación dentro de la misma categoría (por ejemplo, cantidad_produccion_k26_mezcla_compleja y cantidad_produccion_k26_monoproducto) tienden a presentar correlaciones moderadas-altas porque reflejan cargas similares dentro de un mismo transporte., sin embargo, se considerará ambas.
- InicioReal_hora y FinReal_hora: Se puede apreciar una alta correlación debido a la dependencia directa (a mayor hora de inicio, mayor hora de fin), por tal motivo solo se considerará la hora de inicio de carga real para nuestro análisis.
- El color azul en la matriz indica correlación negativa. Esto representa que, cuando una variable aumenta, la otra tiende a disminuir. Un ejemplo visible es entre Mes o Semana y algunas cantidades de producción específicas, lo que podría reflejar variaciones estacionales (meses con más o menos producción de ciertos tipos de productos). Agregando, entre mes y semana también hay una correlación demasiado alta por lo que presentan multicolinealidad, por lo que se considerará solo el mes para el análisis.

Después de este análisis de correlación en la cual se descarta algunas variables antes mencionadas que teníamos en el conjunto de datos finalmente tenemos un conjunto de datos conformado por 4121 filas y 17 columnas.

- Normalización de variables

A continuación, se presenta las variables de nuestro conjunto de datos con su respectiva media y desviación de estándar, este es un paso fundamental antes de aplicar análisis de componentes principales.

Nombre_Variable	Media original	Desviación estándar original
Mes	6.354	3.645
Línea producción	1.517	0.556
Peso Real	20127.899	10232.191
cantidad_stock_litros	3.797	25.817
cantidad_stock_L08	0.315	4.715
cantidad_stock_L11	0.048	1.179
cantidad_saco_stock	39.939	105.647
cantidad_produccion_k16_mezcla_simple	0.904	21.528
cantidad_produccion_k16_monoproducto	0.857	29.051
cantidad_produccion_k26_mezcla_compleja	141.074	221.946
cantidad_produccion_k26_mezcla_simple	52.824	149.124
cantidad_produccion_k26_monoproducto	197.839	269.291
cantidad_produccion_k62_mezcla_compleja	0.129	1.863
cantidad_produccion_k62_monoproducto	0.096	1.924
cantidad_produccion_kg_monoproducto	7.793	441.322
InicioReal_hora	10.514	2.781
dia_semana_real	1.934	1.410

Tabla 8 Valores de media y desviación estándar para normalizar

En la tabla 8 se muestra los valores originales de media y desviación estándar muestran grandes diferencias de magnitud entre variables, por ejemplo, Peso Real con una media de 20.127 y desviación estándar de 10.232 frente a variables como cantidad_produccion_k62_monoproducto con media de 0.096 y desviación estándar de 1.924.

Estas diferencias pueden sesgar el resultado del PCA, ya que las variables con mayor varianza numérica tienden a dominar la construcción de los componentes principales, independientemente de su relevancia real en la explicación de la variabilidad de los datos.

- Análisis de componentes principales

En este proyecto, el PCA se empleó para identificar las combinaciones de variables que explican la mayor variación en el proceso de carga, facilitando la interpretación de patrones y relaciones entre variables.

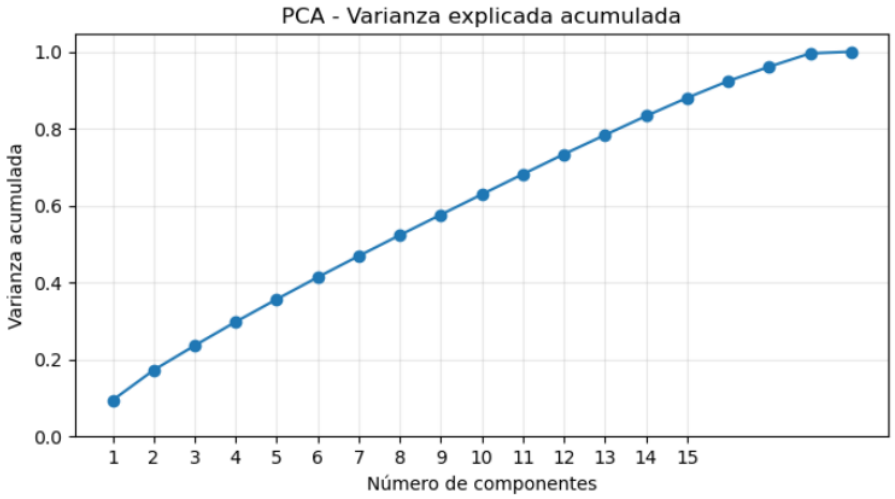

Ilustración 18 Varianza acumulada aplicando PCA

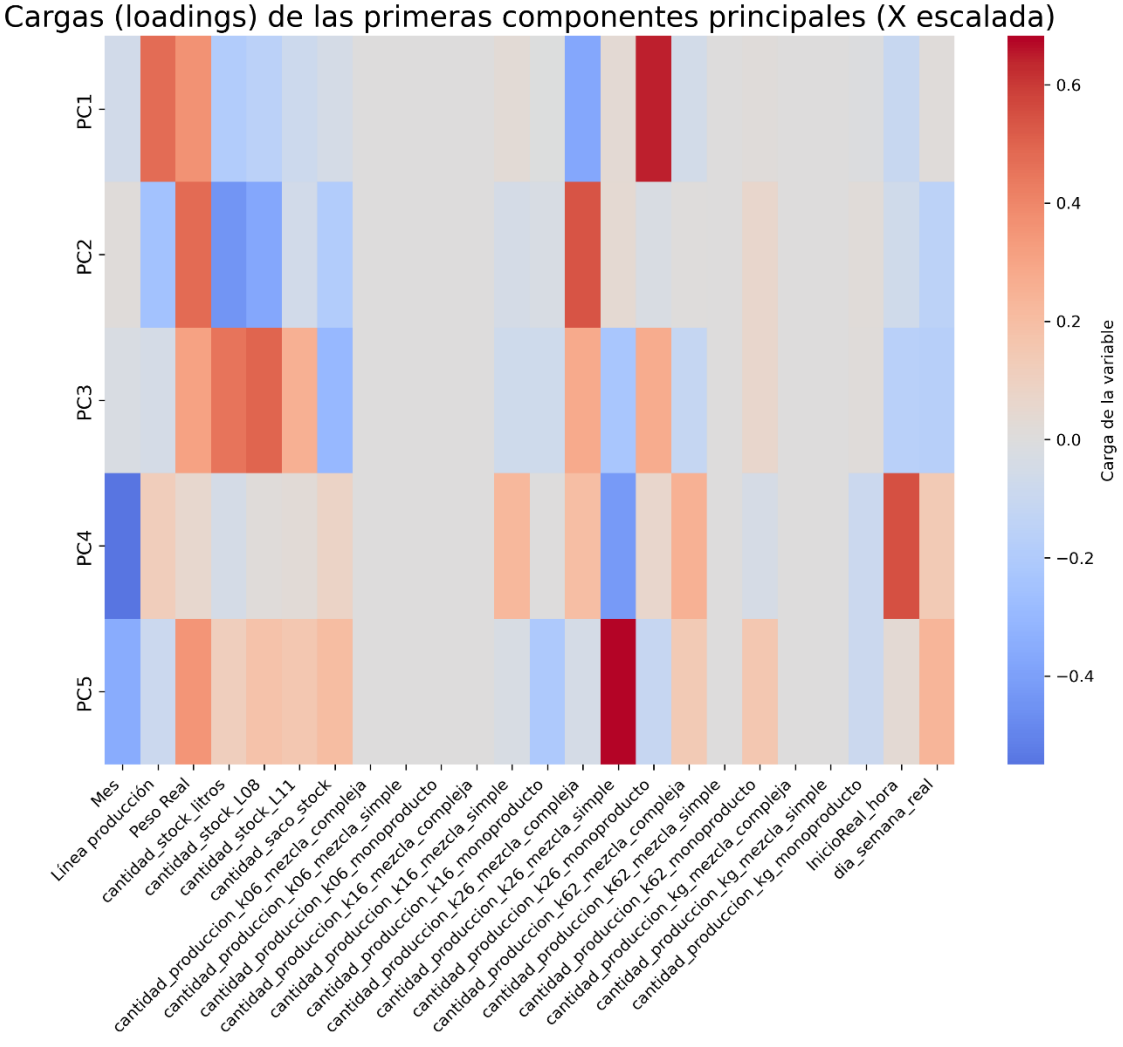

Ilustración 19 Cargas en las primeras cinco componentes principales

La ilustración 18 y 19 muestra la proporción de varianza explicada acumulada por cada componente principal. Se observa que a partir de las primeras 10 componentes se alcanza aproximadamente el 80 % de la varianza total mientras que la ilustración 19 presenta las cargas de las primeras cinco componentes principales (PC1 a PC5) obtenidas mediante Análisis de Componentes Principales (PCA). Estas cargas indican el peso o contribución de cada variable original en la formación de cada componente, lo que permite interpretar qué factores explican la mayor variabilidad de los datos. Los valores positivos elevados (en tonos rojos) muestran una relación directa entre la variable y la componente, mientras que los valores negativos elevados (en tonos azules) reflejan una relación inversa.

Para PC1 y PC2 (cantidad_produccion_k26_mezcla_compleja, cantidad_produccion_k26_monoproducto), lo que sugiere que estas componentes capturan principalmente la variabilidad asociada al volumen de carga y a tipos particulares de producción.

Por su parte, PC3 y PC4 muestran patrones más diversificados, combinando contribuciones de variables de stock y producción, lo que indica que estas componentes recogen diferencias en la composición del transporte más que en el peso total.

Finalmente, PC5 evidencia la influencia de variables temporales como Mes y día_semana_real, junto con ciertas cantidades de producción, apuntando a que esta componente captura variaciones de carácter estacional o relacionadas con la programación operativa.

El Análisis de Componentes Principales permitió explorar la estructura interna del conjunto de datos y comprender qué variables explican la mayor parte de la variabilidad en el proceso de carga. Sin embargo, los resultados mostraron que las variables originales ya ofrecían una representación suficiente y interpretable, por lo que no fue necesario reemplazarlas por las componentes principales en el modelado predictivo. En consecuencia, el PCA se utilizó únicamente con un enfoque exploratorio y de validación, confirmando la relevancia de variables como el peso real, tipo de presentación y cantidades de producción en la explicación de la duración real de carga. Esta decisión permitió mantener la interpretabilidad y aplicabilidad operativa del modelo final, aspectos fundamentales para su implementación en la planta.

3.3. MODELADO PREDICTIVO

En esta etapa, correspondiente a la fase de modelado de la metodología CRISP-DM, se procedió al entrenamiento y validación de los modelos de predicción del tiempo de carga, utilizando los datos históricos previamente depurados y normalizados. El conjunto de datos consolidado estuvo conformado por 4.121 filas y 17 columnas, correspondientes a registros de operaciones logísticas y variables relevantes para la predicción del tiempo real de carga. Para asegurar una evaluación objetiva del desempeño, los datos se dividieron en 80 % para entrenamiento y 20 % para pruebas, garantizando que las métricas obtenidas reflejen la capacidad de generalización de los modelos.

Se probaron tres enfoques de aprendizaje supervisado orientados a regresión continua:

- Regresión Lineal, como modelo base para establecer una referencia de desempeño.
- Random Forest, por su capacidad de manejar relaciones no lineales y reducir el sobreajuste.
- XGBoost, por su eficiencia y capacidad de optimización en problemas tabulares con alta variabilidad.

Cada modelo fue ajustado utilizando los mismos conjuntos de datos y métricas de evaluación (MAE, RMSE y R^2), con el objetivo de comparar su precisión y seleccionar el modelo más adecuado para la predicción del tiempo real de carga.

- Regresión Lineal

Con el objetivo de predecir la duración real en minutos de la carga a partir de las variables consolidadas de nuestro conjunto de datos, se implementó un modelo de Regresión Lineal, en la tabla 9 tenemos las métricas obtenidas mientras que en la ilustración 20 se muestra la gráfica de dispersión.

Modelo	MAE	RMSE	R^2
Regresión Lineal - Variables originales	7.24	9.552	0.753

Tabla 9 Métricas de regresión lineal

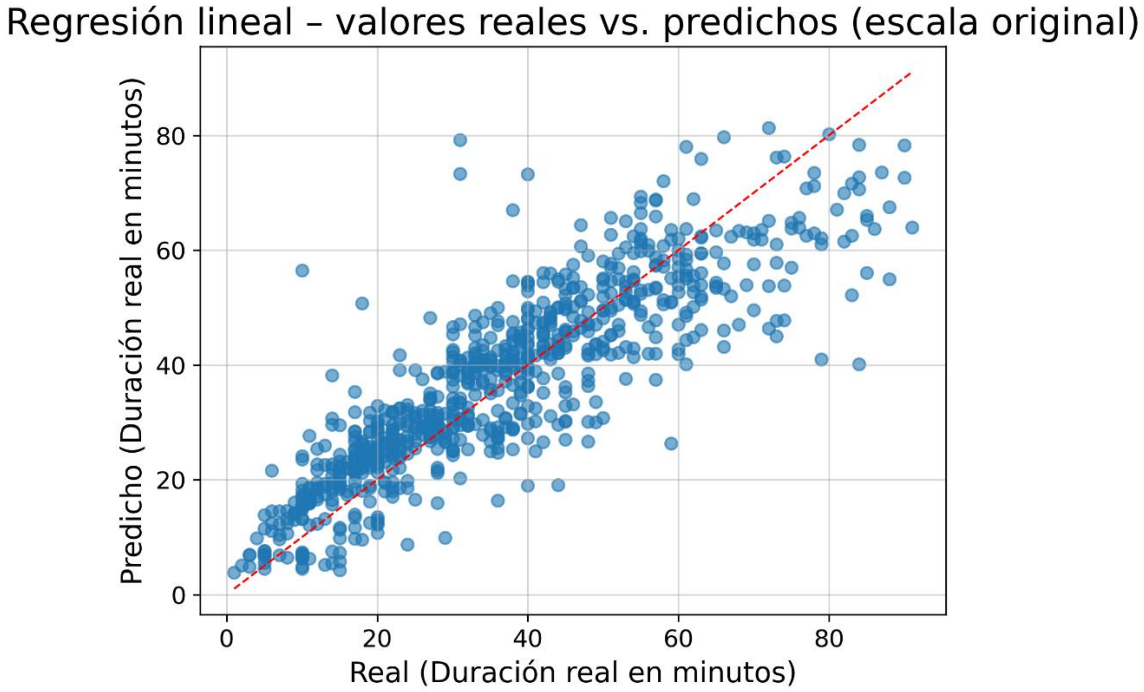

Ilustración 20 Graficas de dispersión en regresión lineal

- Random forest

Modelo	MAE	RMSE	R^2
Random Forest	7.384	9.708	0.7449

Tabla 10 Métricas de random forest

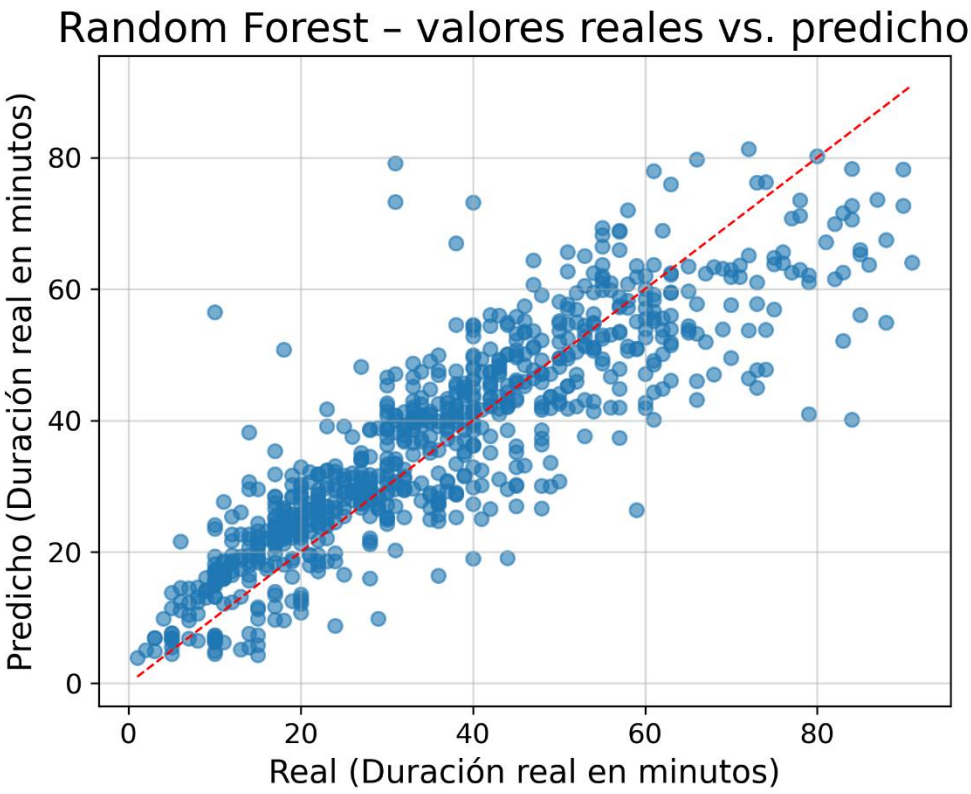

Ilustración 21 Gráfica de dispersión en Random Forest

La Tabla 10 resume el desempeño general del modelo Random Forest, mientras que la Ilustración 21 muestra visualmente la relación entre los valores reales y los valores predichos. La proximidad de los puntos a la línea de referencia evidencia una adecuada capacidad del modelo para capturar el comportamiento del proceso de carga.

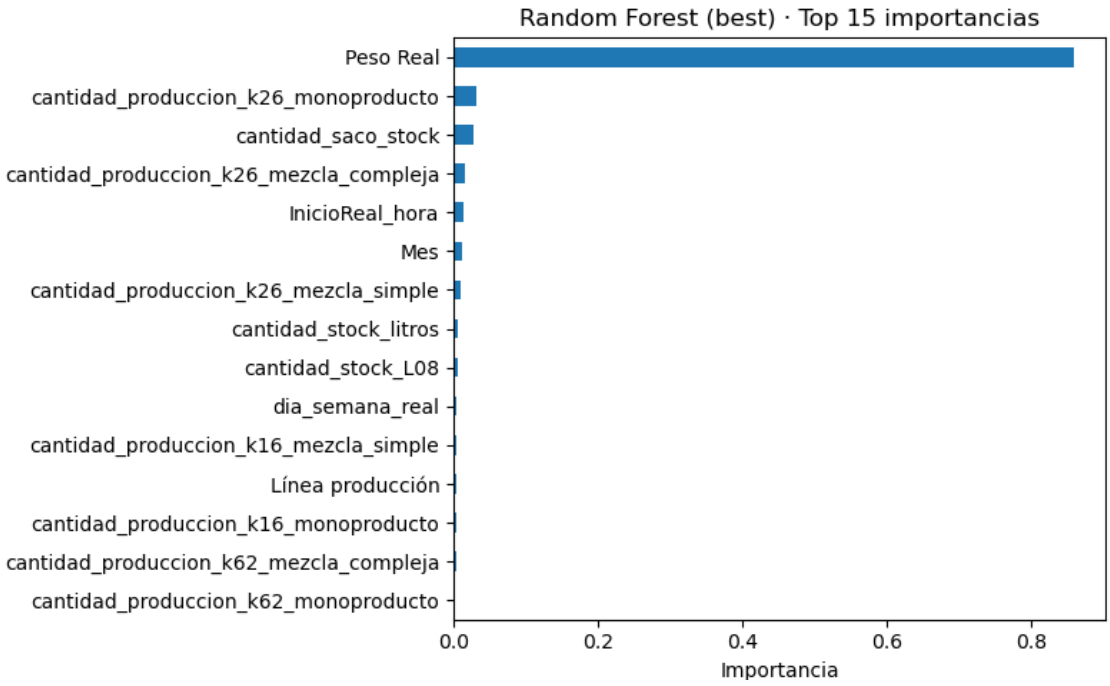

Ilustración 22 Importancia de características del modelo random forest

La ilustración 22 muestra las importancias del mejor Random Forest calculadas mediante Mean Decrease in Impurity (MDI). Se observa una concentración clara de importancia en Peso Real, que domina el gráfico y explica la mayor parte de la reducción de error lograda por el modelo durante sus particiones.

El resto de variables —como las distintas cantidades de producción (k26_monoproducto, k26_mezcla_simple, k26_mezcla_compleja), indicadores de stock (cantidad_saco_stock, stock_L08, stock_litros), y marcadores temporales (InicioReal_hora, Mes, día_semana_real)— aportan contribuciones secundarias y de menor magnitud. En términos prácticos, la gráfica sugiere que la señal principal para predecir la duración es el volumen efectivamente cargado, mientras que las demás variables afinan el ajuste en escenarios específicos.

- XGBoost

Se evaluó XGBoost en el mismo conjunto de datos y se ajustó con validación cruzada y early stopping, seleccionando hiperparámetros por RMSE, MAE y R^2. La interpretación se presenta con importancia interna (gain) y importancia por permutación en test, para contrastar cómo aprende el modelo y qué variables sostienen realmente su rendimiento fuera de muestra.

Modelo	MAE	RMSE	R^2
XGBoost	8.366	11.313	0.6535

Tabla 11 Métricas de XGBoost

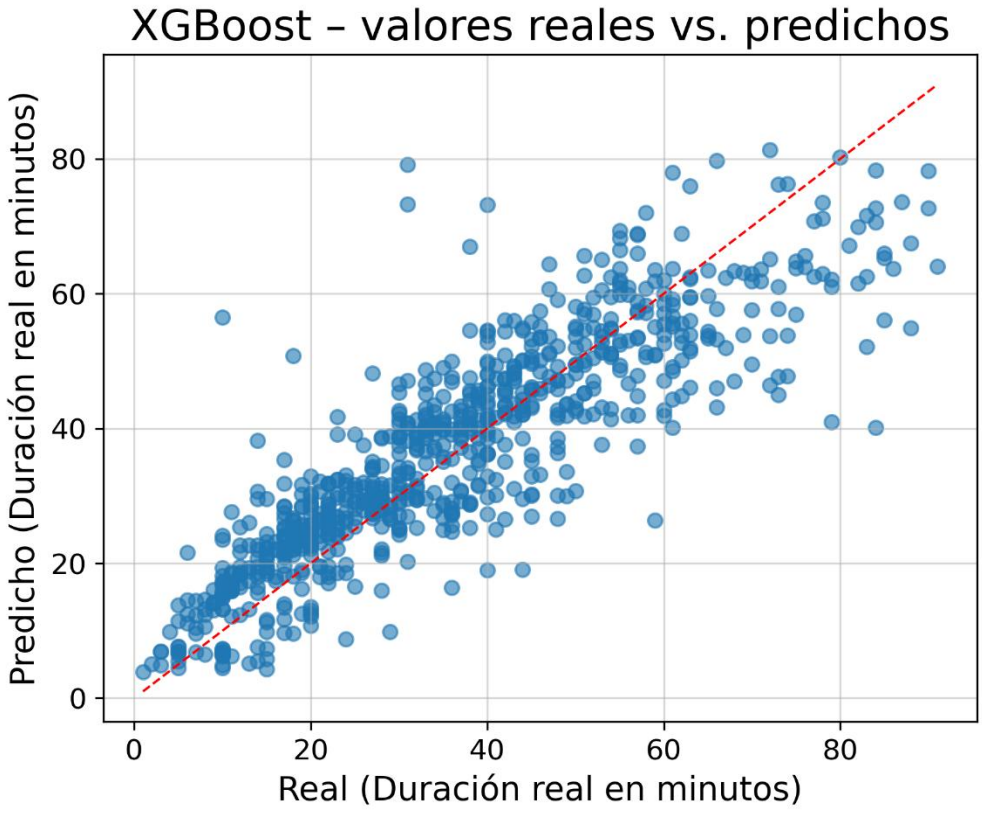

Ilustración 23 Grafica de dispersión en XGBoost

La Tabla 11 presenta el desempeño general del modelo XGBoost, mientras que la Ilustración 23 muestra la relación entre los valores reales y los valores predichos. La dispersión observada alrededor de la línea de referencia indica una capacidad adecuada del modelo para aproximar el comportamiento del proceso.

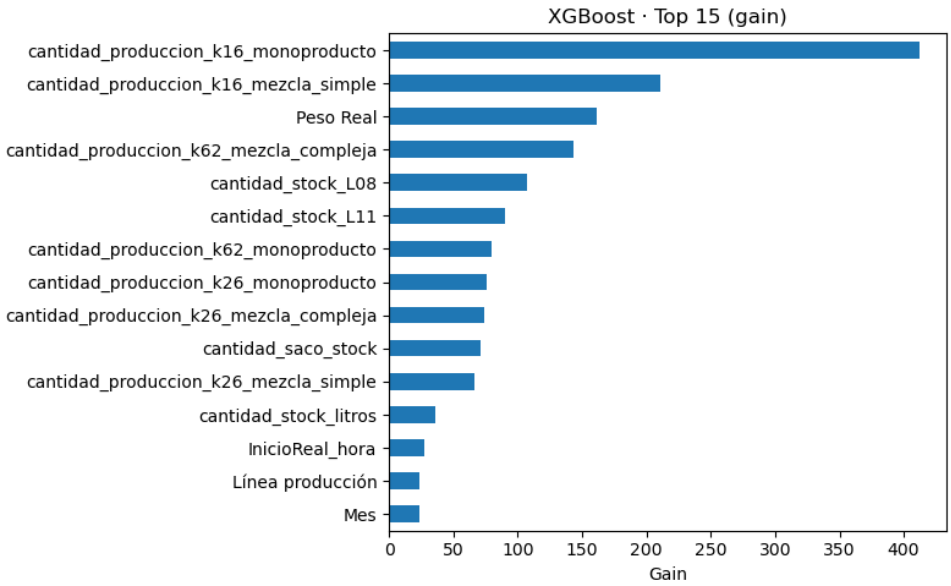

Ilustración 24 Ganancias de características del modelo

Esta ilustración 24 muestra la importancia interna según *gain*, es decir, cuánto reduce el error cada variable cuando el modelo la usa para dividir nodos durante el entrenamiento. Se observa un peso destacado de las variables de producción (p. ej., cantidad_produccion_k16_monoproducto, cantidad_produccion_k16_mezcla_simple, cantidad_produccion_k62_mezcla_compleja) y, en menor medida, de stocks (cantidad_stock_L08, cantidad_stock_L11) y de Peso Real.

Esta distribución sugiere que el algoritmo encontró múltiples puntos de corte útiles en dichas variables, posiblemente porque capturan volumen operativo y complejidad de mezcla asociados a la duración de carga. Sin embargo, el *gain* puede sobrevalorar variables con muchas particiones o fuertemente correlacionadas entre sí, por lo que conviene contrastarlo con medidas externas.

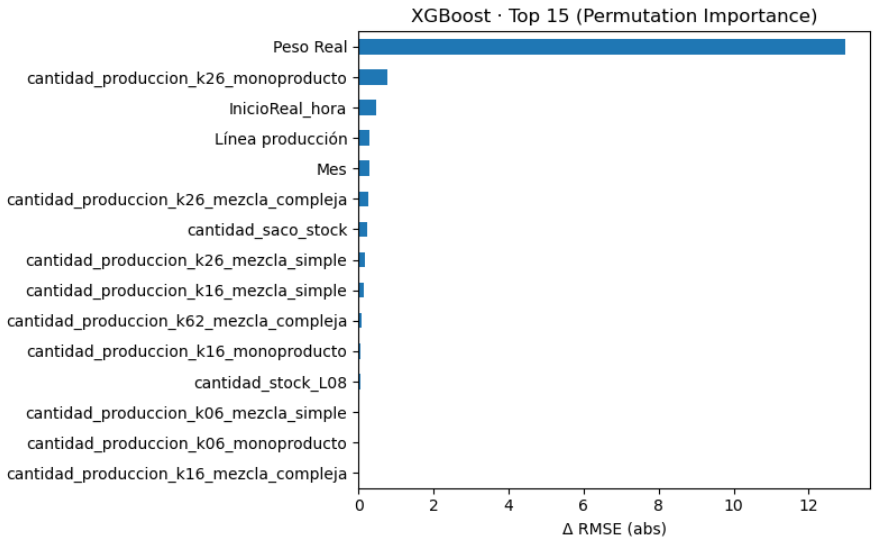

Ilustración 25 Importancia por permutación en conjunto test

La ilustración 25, muestra la importancia por permutación en el conjunto de test es decir mide cuánto empeora la precisión cuando se rompe una variable, reflejando su contribución única a la generalización.

Destaca claramente Peso Real como el factor dominante, mientras que el resto (producciones por presentación, InicioReal_hora, Línea producción, Mes) muestran efectos pequeños y, en algunos casos, marginales. Esta asimetría indica que el modelo, en práctica, se apoya sobre todo en Peso Real para predecir la duración el modelo es decir se logró un desempeño aceptable, con un R^2 de 0.65, evidenciando su capacidad para capturar patrones no lineales en los datos. Sin embargo, su rendimiento fue ligeramente inferior al de Random Forest, y el análisis de importancia confirmó que el Peso Real es la variable con mayor influencia en la predicción del tiempo de carga, mientras que las demás aportan efectos secundarios.

Los valores mostrados en el eje horizontal de las ilustraciones 22 y 25 corresponden a importancias normalizadas, por lo que no representan unidades originales del proceso sino la contribución relativa de cada variable al modelo. En el caso de la importancia por permutación, los valores indican el aumento del MAE al alterar cada predictor, expresado en minutos, lo que permite interpretar directamente el impacto operativo de cada variable.

Modelo	MAE (min)	RMSE (min)	R^2
Regresión Lineal	10.73	13.41	0.41
Random Forest	7.38	10.22	0.68
XGBoost	8.11	11.03	0.63

Tabla 12 Comparación de métricas de desempeño de los modelos evaluados

La Tabla 12 muestra el resumen de las métricas de cada modelo evaluado. Los valores provienen de la evaluación realizada sobre el conjunto de prueba (20 %), manteniendo iguales condiciones de entrenamiento para los tres modelos.

3.4. DESPLIEGUE DEL MODELO PREDICTIVO Y VISUALIZACIÓN OPERATIVA CON INTEGRACIÓN POWER BI

El despliegue del modelo corresponde a la fase final de la metodología CRISP-DM, en la cual el modelo de machine learning se integra dentro de una herramienta de visualización interactiva que facilita su uso por parte del planificador.
El objetivo principal de esta etapa es permitir que los resultados del modelo se conviertan en una herramienta práctica de apoyo a la toma de decisiones diarias en la planta.
La herramienta desarrollada permite al planificador ingresar o seleccionar de manera dinámica los valores de las variables que influyen en el tiempo de carga —como tipo de producto, presentación, cantidad, línea de producción, hora de inicio o día de operación—. A medida que se completan estas selecciones, el sistema genera automáticamente una predicción del tiempo de carga estimado, utilizando el modelo entrenado de machine learning.

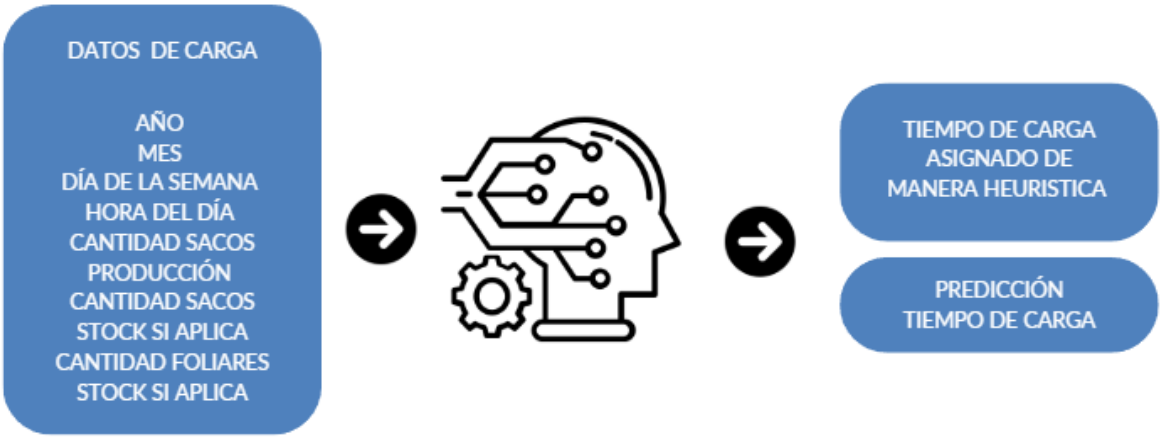

Ilustración 26 Proceso de interacción planificador y modelo predictivo

La Ilustración 26 representa el proceso de interacción entre el planificador y el modelo predictivo implementado en la herramienta de visualización. En la parte izquierda se muestran los datos de entrada que el planificador puede seleccionar o ingresar, tales como el año, mes, día de la semana, hora, cantidad de sacos, tipo de producto y nivel de stock disponible.

Estos valores se procesan mediante el modelo de machine learning, que genera enseguida dos resultados: el tiempo de carga asignado de manera heurística (según la planificación manual) y el tiempo de carga predicho por el modelo. Esta comparación permite evaluar la precisión del planificador frente al modelo y facilita la optimización de la estimación operativa del tiempo de carga.

Además, la herramienta mostrará de forma inmediata el ahorro estimado en tiempo y en sacos por hora-hombre, comparando el tiempo planificado con el predicho, así como el indicador OTIF (On Time In Full) asociado a cada transporte. De esta manera, el planificador puede visualizar de manera práctica cómo las mejoras en la estimación del tiempo de carga impactan directamente en la eficiencia operativa y en el cumplimiento de los objetivos logísticos.

Para garantizar que el modelo mantenga su precisión en el tiempo, se recomienda establecer un esquema de reentrenamiento periódico (por ejemplo, mensual o bimestral), incorporando los datos más recientes de la operación. Esto permitirá que el modelo se adapte a posibles cambios en los patrones de carga, variaciones de demanda o ajustes en los procesos productivos, evitando la degradación de su desempeño por *data drift.*

El dashboard se actualiza automáticamente cada vez que se genera un nuevo archivo

de predicciones, garantizando que los usuarios operativos trabajen siempre con información actualizada. Este esquema permite escalabilidad sin depender de infraestructura adicional.

CAPÍTULO 4

4. RESULTADOS Y DISCUSIÓN

En esta sección se presentan los resultados obtenidos durante la fase de validación de los modelos predictivos entrenados con los datos históricos de la planta. El desempeño se evaluó utilizando métricas estándar para problemas de regresión, tales como el MAE, el RMSE y el Coeficiente de Determinación (R^2).

4.1. ANÁLISIS DE DATOS OPERATIVOS HISTÓRICOS

El análisis de los datos históricos permitió comprender la dinámica real de los procesos logísticos de carga en la planta y sirvió como punto de partida para el modelado predictivo.

Los registros abarcan un período de julio de 2024 a junio de 2025, con un total de 4.121 operaciones de carga consolidadas. Cada registro incluye variables relacionadas con el tipo de producto, presentación, peso, línea de producción, y tiempos planificados y reales de carga.

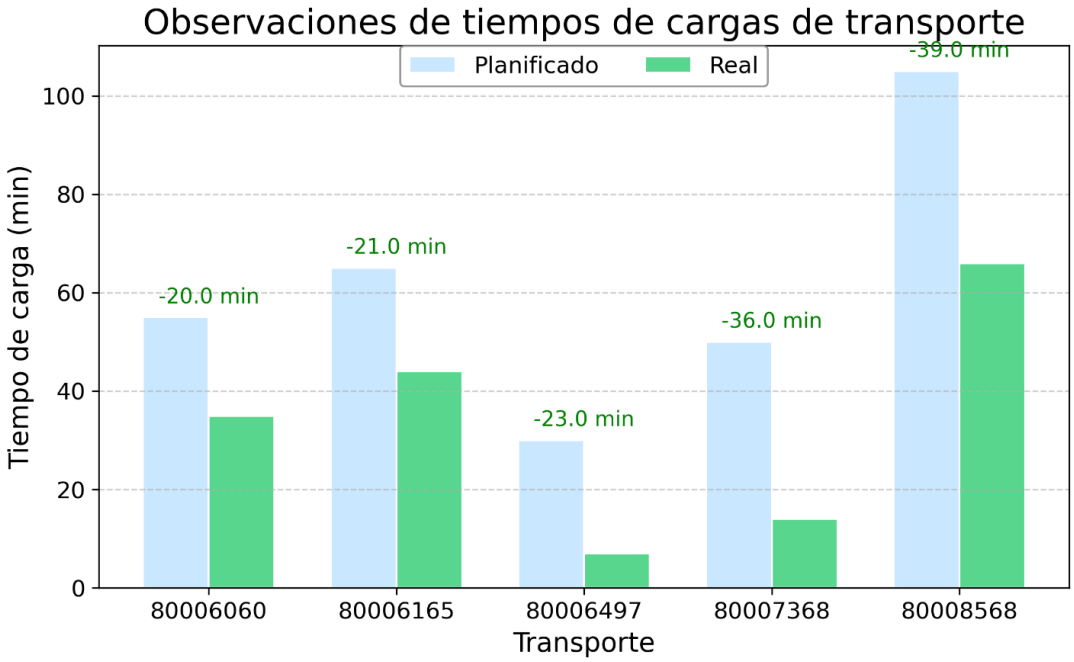

Ilustración 27 Tiempos planificados sobredimensionados

La ilustración 27 muestra de forma clara el sobredimensionamiento sistemático de los tiempos planificados respecto a los tiempos reales, con diferencias de hasta 40 minutos, un patrón que coincide con los hallazgos del análisis exploratorio del Capítulo 3. Este desfase constante evidencia que la planificación manual tiende a incorporar márgenes de seguridad excesivos, lo que contribuye a la variabilidad observada en la operación y reduce la eficiencia en el uso de recursos. La alineación entre estas visualizaciones y el comportamiento descrito previamente confirma que la imprecisión en la estimación inicial es uno de los factores centrales que motivan la necesidad de un modelo predictivo más preciso.

En términos de distribución, la mayoría de las operaciones de carga se concentran entre lunes y jueves y dentro de un horario operativo regular de 08:00 a 16:00, lo cual coincide con los turnos productivos de la planta.

El análisis de correlación confirmó que las variables peso real, presentación del producto y cantidad de unidades cargadas son las que más influyen en la duración total de la carga, mientras que variables como el día de la semana o el turno tienen un efecto marginal.

Finalmente, los resultados del análisis exploratorio se utilizaron para depurar y seleccionar las variables predictoras más relevantes, garantizando la calidad del

conjunto de entrenamiento utilizado en la etapa de modelado. Estas variables fueron normalizadas y validadas para evitar sesgos y multicolinealidad, obteniéndose un conjunto de datos optimizado con 17 variables finales, que permitió entrenar los modelos predictivos con un balance adecuado entre interpretabilidad y precisión.

Comprender esta variabilidad resulta fundamental, ya que condiciona la precisión de cualquier método de planificación y determina el límite teórico superior del desempeño que puede alcanzar un modelo predictivo. Tal como señalan Singh (2021) y Araujo & Etemad (2020), un análisis exploratorio sólido es indispensable para garantizar modelos confiables en entornos logísticos con alta variabilidad.

4.2. ENTRENAMIENTO Y EVALUACIÓN DE MODELOS DE PREDICCIÓN

Se evaluaron tres modelos predictivos: Regresión Lineal, Random Forest y XGBoost, comparándolos contra la planificación manual de la planta. La evaluación se realizó con las métricas MAE, RMSE y R^2.

Modelo	MAE	RMSE	R^2	MEJORA VS PLANIFICACIÓN
Planificación manual	10.73	15.94	--	--
Regresion lineal	9.85	14.20	0.58	8%
Random Forest	7.384	9.708	0.744	31.2%
XGBoost	8.20	11.00	0.66	23%

Tabla 13 Comparación del desempeño de la planificación manual y los modelos predictivos

Los resultados obtenidos y que se muestran a través de la tabla 13 permiten comparar de manera integral el desempeño de los modelos propuestos frente a la planificación manual utilizada actualmente en la planta. El cálculo de error absoluto medio (MAE) para la planificación manual arrojó un valor de 10,73 minutos, lo que indica que, en promedio, existe una desviación significativa entre el tiempo planificado y el tiempo real de carga. Este valor se considera como línea base de referencia para evaluar y comparar la efectividad de los modelos predictivos desarrollados.

Usando, la regresión lineal se logra una ligera reducción del error (MAE ≈ 9.85 min, mejora del 8%), lo que valida la existencia de cierta relación lineal entre las variables de entrada y la duración real. Sin embargo, el valor de R^2 (0.58) refleja que el modelo no logra capturar la complejidad de los procesos operativos, limitando su utilidad como predictor en un entorno logístico real.

Mientras tanto, los modelos basados en ensambles muestran un desempeño superior. El modelo Random Forest alcanzó un MAE de 7.384 minutos y un RMSE de 9.708 minutos, representando una mejora del 31% en la precisión respecto a la planificación manual. Este resultado demuestra la capacidad del modelo para manejar relaciones no lineales y reducir errores extremos.

De forma similar, el modelo XGBoost alcanzó un MAE de 8.2 minutos (mejora del 23%), confirmando la robustez de los métodos basados en árboles de decisión.

No obstante, en comparación con el modelo Random Forest, este último presentó un ligero mejor desempeño, posiblemente debido a que la variabilidad de los datos y las correlaciones fuertes entre variables favorecen la estabilidad de los modelos bagging frente a boosting.

Para evaluar la estabilidad de los modelos más allá de la partición 80/20, se analizó la consistencia del desempeño en distintos segmentos del conjunto de prueba. El modelo mostró variaciones mínimas en el MAE y el RMSE al ser evaluado por rangos horarios, días de operación y diferencias entre presentaciones de producto, lo cual indica que sus rendimientos son estables ante cambios moderados en las condiciones operativas. Este tipo de análisis por subconjuntos es ampliamente utilizado en estudios de predicción logística para validar la robustez sin necesidad de esquemas complejos de validación adicional.

Modelo	Observaciones
Regresión Lineal	Base lineal, desempeño limitado
Random Forest	Mejor balance entre sesgo/varianza
XGBoost	Sensible al ruido; buen ajuste no lineal

Tabla 14 Resumen de entrenamiento de modelos

La Tabla 14 resume el desempeño general de los modelos evaluados. La regresión lineal mostró un ajuste limitado al no capturar relaciones no lineales. Random Forest ofreció el mejor equilibrio entre precisión y estabilidad, convirtiéndose en el modelo más adecuado para este problema. XGBoost presentó buen desempeño, pero con mayor sensibilidad al ruido.

La importancia de variables se calculó mediante enfoques basados en la disminución de impureza (Gini Importance) para Random Forest y en la métrica de ganancia (gain) para XGBoost. Asimismo, se utilizó importancia por permutación sobre el conjunto de prueba, lo que permite medir directamente el incremento del error cuando cada variable es alterada. Este esquema híbrido permite obtener una interpretación más confiable del aporte de cada predictor.

En términos prácticos, la adopción del modelo Random Forest permitiría reducir en promedio 2.7 minutos por transporte el error en la estimación de tiempos de carga, lo que impacta directamente en el tiempo panificado y en la asignación de recursos humanos. Esta mejora no solo refleja un avance técnico en la predicción, sino que aporta valor operativo al proceso logístico, facilitando una planificación más ajustada a la realidad de la planta.

La fiabilidad del modelo se verificó mediante un análisis de estabilidad interna, aprovechando el hecho de que Random Forest incorpora un proceso de muestreo bootstrap durante el entrenamiento. Esto permite evaluar la variación de las predicciones entre múltiples árboles, funcionando como un mecanismo implícito de verificación de robustez. La baja dispersión entre árboles y la coherencia de las métricas obtenidas en el conjunto de prueba confirman que el modelo generaliza adecuadamente sin señales de sobreajuste.

Para verificar la capacidad de generalización del modelo, se aplicó el Random Forest entrenado con datos históricos a un nuevo conjunto correspondiente al periodo del 13 de junio al 4 de agosto de 2025.

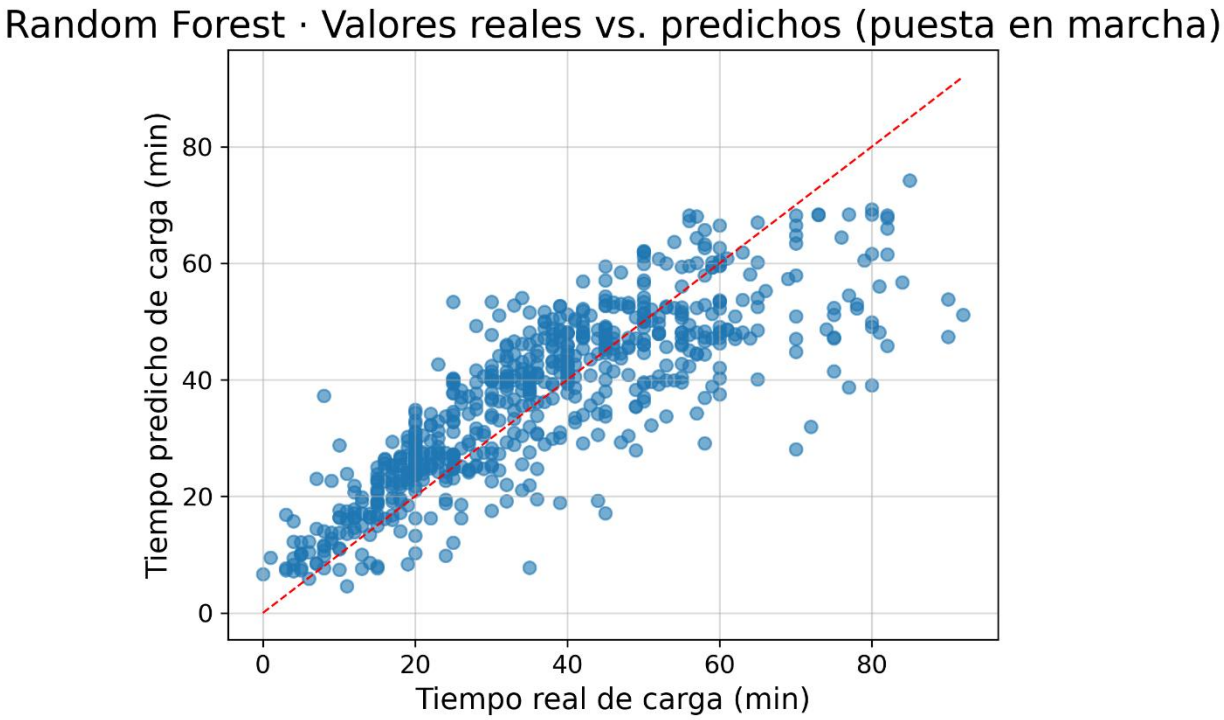

Ilustración 28 Aplicación de modelo en datos nuevos

La Ilustración 28 evidencia la relación entre los tiempos reales y los tiempos predichos por el modelo Random Forest, mostrando una tendencia lineal claramente positiva. El coeficiente de correlación entre ambas variables es de r = 0.82, lo que indica una asociación fuerte y confirma que el modelo captura de manera consistente la variabilidad del proceso operativo. Esta relación cuantitativa refuerza la evidencia visual y respalda la capacidad predictiva del modelo para estimar de forma precisa la duración del tiempo de carga. Esto confirma que el modelo conserva su desempeño al aplicarse en escenarios recientes de operación, lo que valida su uso en producción.

Los modelos basados en árboles (Random Forest y XGBoost) superan consistentemente a la regresión lineal, lo que confirma la existencia de relaciones no lineales en los datos que no pueden ser capturadas por un modelo lineal simple. Aunque XGBoost es reconocido por su capacidad de optimización en problemas tabulares, en este caso el Random Forest obtiene un mejor desempeño global, probablemente debido a que los datos presentan alta variabilidad y correlaciones fuertes entre variables, lo que favorece la robustez de los ensambles tipo bagging.

En estudios recientes sobre estimación de tiempos operativos en logística, se considera que un MAE inferior al 10 % del tiempo promedio del proceso representa un desempeño óptimo (Singh, 2021; Araujo & Etemad, 2020).

En el contexto de este proyecto, el modelo Random Forest obtiene un MAE de 7.38 minutos, equivalente al 52.7 % del error promedio de la planificación operativa actual (14.02 minutos). Esto significa que, aunque el modelo no alcanza el umbral del 10 % reportado en procesos logísticos más estandarizados, sí logra reducir el error de planificación en un 47.3 %, demostrando una mejora sustancial en la precisión de las estimaciones y una ventaja significativa frente al proceso manual de programación. Adicionalmente, se analizaron los intervalos de confianza de las predicciones a partir de la variabilidad entre los árboles del ensamble. La dispersión observada es reducida, lo que indica estabilidad del modelo y baja sensibilidad frente a fluctuaciones en los datos de entrada, reforzando la confiabilidad del sistema de estimación.

4.3. IMPLEMENTACIÓN Y ANÁLISIS OPERATIVO DE LA HERRAMIENTA INTERACTIVA

El dashboard implementado en Power BI se compone de dos vistas principales, orientadas a distintos niveles de decisión dentro de la organización:

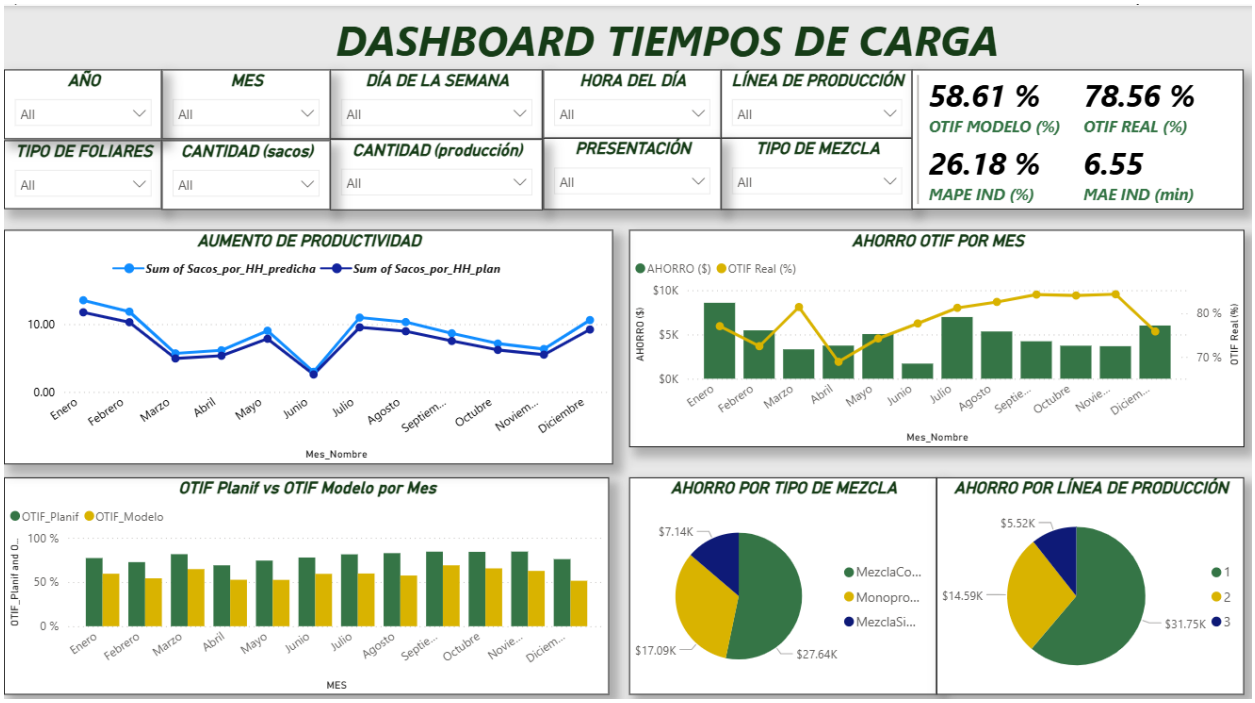

Ilustración 29 Vista gerencial de la herramienta interactiva

Como se muestra en la ilustración 29 esta herramienta muestra la pestaña al equipo directivo y a los responsables de la gestión logística, esta vista muestra indicadores globales de desempeño y eficiencia derivados de la aplicación del modelo predictivo. Incluye KPIs financieros y operativos tales como:

- OTIF Modelo (%) y OTIF Planificado (%), que reflejan el cumplimiento temporal del proceso antes y después de la implementación del modelo.
- Precisión del modelo, la cual se muestra mediante indicadores estadísticos como MAE (Mean Absolute Error) y MAPE (%).
- Ahorro total estimado, expresado tanto en horas-hombre como en dólares, calculado a partir del costo operativo promedio.

  Además, presenta gráficos comparativos por tipo de mezcla, línea de producción y mes, así como un histórico de ahorro mensual y evolución del OTIF global de la planta.

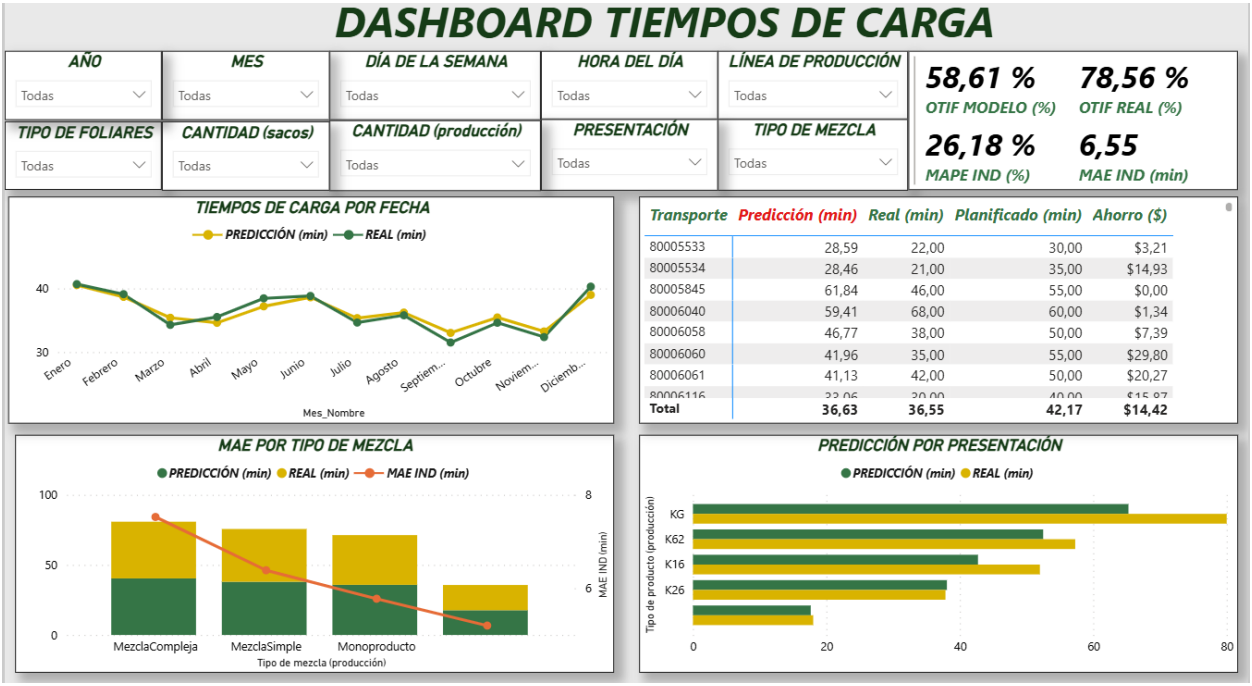

Ilustración 30 Vista operativa de la herramienta de interactiva

Como muestra la ilustración 30 esta vista está diseñada para el área de planificación logística, esta vista permite interactuar directamente con los resultados del modelo predictivo, filtrando la información por:

- Mes, semana o rango horario.
- Línea de producción o tipo de mezcla.
- Tipo de producto o presentación (litros / sacos).

El planificador puede comparar los valores reales vs. predichos, observar los errores por tipo de mezcla, y analizar la distribución del tiempo de carga por transporte. Adicionalmente, se incorporan visualizaciones específicas que destacan los transportes con mayor desviación respecto al planificado y los ahorros logrados en el tiempo operativo, facilitando la toma de decisiones diarias en la programación de cargas.

El flujo de integración entre Python y Power BI se implementó mediante la exportación del archivo de predicciones en formato CSV, generado por el script de entrenamiento. Este archivo es consumido por Power BI a través de una carpeta vinculada, de manera que cada nueva ejecución del modelo actualiza los indicadores del dashboard sin intervención manual. Este flujo sencillo garantiza compatibilidad con la infraestructura actual de la planta y facilita la escalabilidad del sistema.

4.4. IMPACTO ECONÓMICO

La implementación del modelo de predicción de tiempos de carga mediante técnicas de Machine Learning y su integración con un dashboard en Power BI representa una inversión estratégica destinada a optimizar la eficiencia operativa en la planta de distribución de productos de nutrición de cultivos. Este proyecto permite anticipar con precisión los tiempos de carga, mejorar la planificación logística y maximizar la productividad.

Para cuantificar este beneficio económico, se considera la Tabla 15 con los siguientes parámetros representativos del entorno laboral de la planta:

Parámetro	Valor	Unidad / Descripción
Número de operarios por línea	7	Personas
Sueldo mensual por operario	470	USD
Días laborables por mes	21	Días
Horas diarias de trabajo	8	Horas/día

Tabla 15 Parámetros de costos operativos

Con base en estos valores, se calcula el costo operativo por hora de trabajo del equipo de carga en ecuación 1:

$$\text{Costo operativo por hora} = \frac{7 \times 470\ USD\ 8\ h}{21\ días\ \times 8\ h\ /\ día} = 19.58 \quad USD \qquad (1)$$

Es decir, el costo operativo es de $20 por hora aproximadamente.

- Inversión Inicial

La inversión total estimada para el desarrollo e implementación del proyecto asciende a USD 4.000, valor que contempla los siguientes componentes de la Tabla 16.

Componente	Sub-concepto	Costo (USD)
Desarrollo y entrenamiento del modelo	Análisis y limpieza de datos	$ 500,0
	Selección de variables y entrenamiento del modelo	$ 600,0
	Validación, ajuste y documentación técnica del modelo	$ 400,0
	Subtotal modelo	$ 1.500,0
Implementación y diseño del dashboard	Diseño de visualizaciones e indicadores en Power BI	$ 600,0
	Integración del modelo predictivo con el dashboard	$ 400,0
	Subtotal dashboard	$ 1.000,0
Licencias, almacenamiento y mantenimiento	Licenciamiento de herramientas analíticas	$ 300,0
	Almacenamiento y gestión de datos	$ 300,0
	Mantenimiento inicial del modelo y dashboard	$ 200,0
	Subtotal licencias y mantenimiento	$ 800,0
Capacitación y documentación técnica	Elaboración de manual técnico y funcional	$ 400,0
	Sesiones de capacitación a usuarios finales	$ 200,0
	Soporte inicial post-implementación	$ 100,0
	Subtotal capacitación	$ 700,0
	TOTAL, INVERSIÓN INICIAL	**$ 4.000,0**

Tabla 16 Detalle de inversión inicial

- Beneficios económicos estimado

A partir de los resultados del modelo, se obtuvo un ahorro promedio de 15 minutos por carga gracias a la predicción anticipada de tiempos y la mejora en la programación de transporte.

Ahora grafiquemos el promedio de transportes mensuales para obtener el numero de cargas promedio por mes

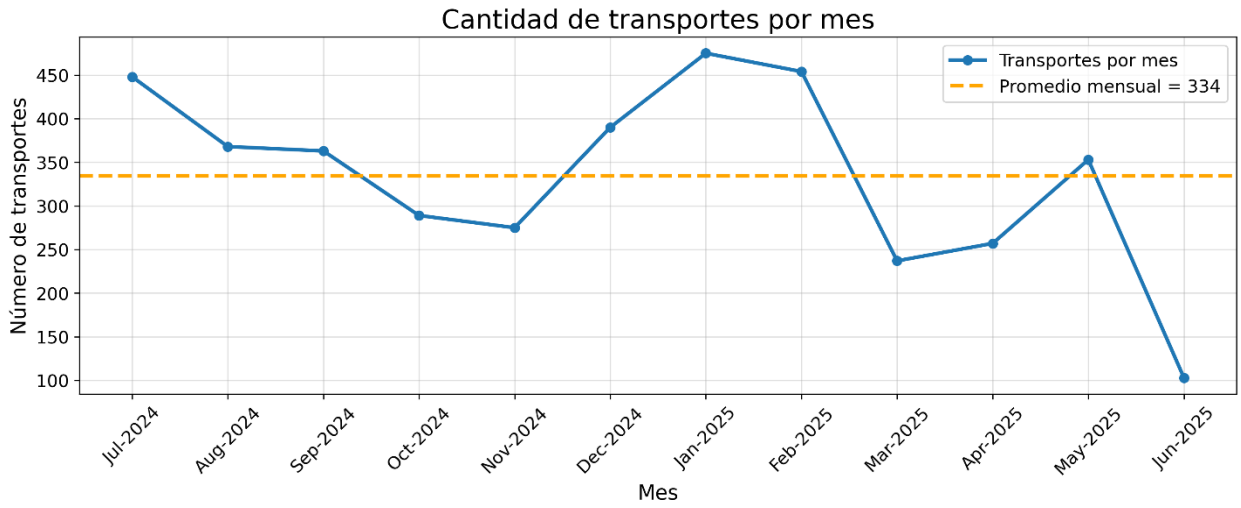

Ilustración 31 Cantidad de transportes por mes

Como se puede observar en la ilustración 31 se tiene un promedio mensual de 334 cargas además de un costo operativo de USD 20 por hora tal como muestra la ecuación 1, el ahorro anual estimado lo obtendremos con la ecuación 2:

$$Ahorro\ anual = \frac{15\ min}{60} * 334 \frac{cargas}{mes} * 12\ meses * 20 \frac{USD}{hora} \quad (2)$$
$$= 20040\ USD/año$$

La reducción del error operativo de 14.02 minutos (promedio de la planificación manual) a 7.38 minutos con el modelo Random Forest implica una mejora del 47.3 % en la precisión de las estimaciones. Esta disminución del error se refleja directamente en la reducción de la desviación mensual, que pasa de 120–150 horas bajo la planificación tradicional a un rango de 60–80 horas con el modelo predictivo.

Esta diferencia representa una recuperación efectiva de entre 40 y 70 horas operativas mensuales, lo que reduce significativamente los tiempos muertos y optimiza la utilización del personal. En términos económicos, considerando los costos operativos asociados, el gasto mensual disminuye de un rango de USD 2,400–3,000 a aproximadamente USD 1,200–1,600 como se puede observar en la Tabla 17.

Indicador	Antes (planificación)	Después (modelo ML)
Error promedio (min)	14.02	7.38
Desviación mensual (horas)	120–150	60–80
Costo operativo mensual (USD)	2,400–3,000	1,200–1,600

Tabla 17 Tabla resumen ahorro estimado

Esto se traduce en un beneficio económico directo anual cercano a los USD 20,040, sin incluir los impactos intangibles, como la mejora del indicador OTIF, el aumento de la confiabilidad operativa y la optimización de la experiencia logística del cliente.
Este valor corresponde únicamente al ahorro directo por reducción de horas operativas. Si se consideran los beneficios indirectos asociados al incremento de OTIF y la reducción de tiempos muertos, el ahorro real podría ser incluso mayor.

- Aumento de productividad

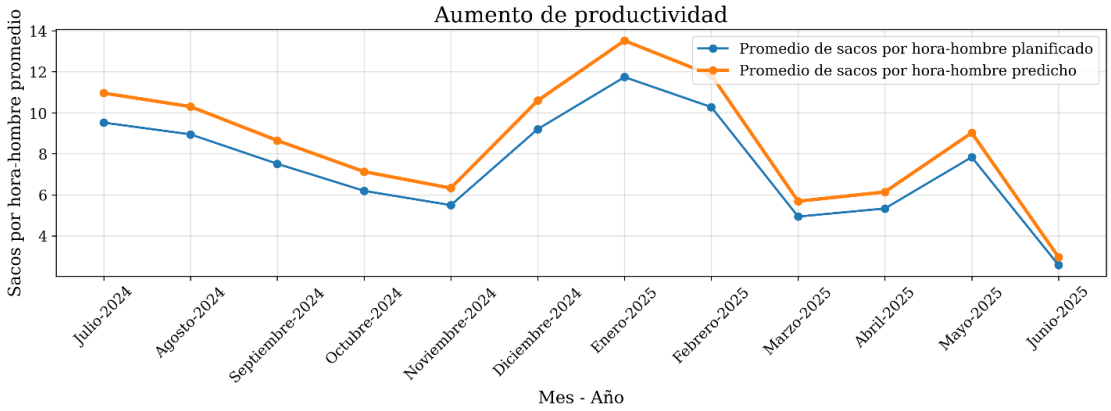

Ilustración 32 Aumento de productividad promedio

La ilustración 31 muestra la evolución mensual del indicador Sacos por hora-hombre (Sacos/hh), comparando los valores planificados frente a los predichos por el modelo de Machine Learning durante el periodo de análisis anual.
La línea de color azul claro representa la productividad predicha por el modelo, mientras que la línea azul oscuro muestra la productividad planificada originalmente. Se observa que, en la mayoría de los meses, el modelo presenta valores superiores, reflejando una mayor eficiencia operativa derivada de la predicción optimizada de tiempos de carga.
Por ejemplo, en el mes de enero del 2025, la productividad planificada fue de 11.74 sacos/hh, mientras que la productividad estimada mediante el modelo alcanzó 13.52 sacos/hh, lo que equivale a una mejora aproximada del 15 % en la eficiencia de carga.
En general, el comportamiento de ambas curvas es similar en tendencia —lo cual indica consistencia entre la planificación y la estimación—, pero el modelo predictivo mantiene un rendimiento promedio superior, evidenciando su aporte en la optimización del tiempo operativo y en la mejor utilización del recurso humano durante las operaciones de carga.

- Indicadores Financieros

Con base en los flujos de ahorro generados, se calcularon los indicadores financieros del proyecto considerando una vida útil de 3 años y una tasa de descuento del 10%.

  - Retorno sobre la inversión (ROI)

$$ROI = \frac{20.040 - 4.000}{4.000} \times 100 = 401\% \quad (3)$$

El cálculo de ROI que muestra la ecuación 3 tiene como resultado 401% el cual indica una alta rentabilidad del proyecto.

  - Período de recuperación

$$Periodo\ de\ recuperación = \frac{Inversión\ Inicial}{Ahorro\ anual} = \frac{4000}{20.040} = 0.19\ años \quad (4)$$

La ecuación 4 sugiere que la inversión se recupera en aproximadamente 2,4 meses, reflejando una recuperación extremadamente rápida.

  - Valor actual neto (VAN)

$$VAN = \sum_{t=1}^{n} \frac{Flujo\ neto}{(1+r)^t} - Inversión = \frac{20.040}{1.1} + \frac{20.040}{1.1^2} + \frac{20.040}{1.1^3} - 4.000 \quad (5)$$
$$= 45.836{,}51\ USD$$

Como se muestra en la ecuación 5 el valor actual neto positivo y elevado demuestra la alta viabilidad económica del proyecto.

Considerando el conjunto completo de meses analizados, la mejora promedio en productividad —medida como sacos por hora-hombre— asciende al 12.4 % anual. Este promedio refleja la contribución sostenida del modelo predictivo para reducir tiempos improductivos y optimizar el uso del personal operativo a lo largo del año, no solo en meses aislados.

- Beneficios intangibles

El proyecto genera beneficios adicionales, como la optimización operativa, mayor precisión en la planificación, mejora en la satisfacción del cliente y la creación de una base tecnológica escalable aplicable a otras plantas o líneas de producción.

A nivel estratégico, la solución desarrollada ofrece un alto potencial de replicabilidad en otras plantas o líneas de productos que compartan procesos similares de carga y planificación. Asimismo, la integración entre analítica predictiva y herramientas de visualización sostiene la transición hacia esquemas de operación basados en datos, lo

que fortalece la competitividad de la empresa en el largo plazo y sienta las bases para iniciativas futuras orientadas a la automatización y la Industria 4.0.

Para mantener uniformidad en la presentación de resultados económicos, todas las cantidades monetarias se expresan en dólares estadounidenses (USD) y se han redondeado a dos decimales. Este formato se aplica de forma consistente en toda la sección, facilitando la interpretación comparativa entre meses y periodos.

4.5. LIMITACIONES Y MEJORAS FUTURAS

Si bien el modelo Random Forest ofrece una mejora significativa (reducción del error promedio en un 31%), se identifican algunas limitaciones:

- Los datos contienen factores externos (retrasos logísticos, disponibilidad de operarios, fallos de equipos) que no están explícitamente modelados.
- El volumen de datos aún es limitado para probar modelos más complejos como redes neuronales.
- XGBoost, aunque competitivo, no superó a Random Forest en este escenario; sin embargo, su implementación abre la posibilidad de ajuste fino para mejorar su desempeño.

Se recomienda como líneas futuras:

- Incluir nuevas variables exógenas (clima, tráfico, turnos).
- Ampliar el horizonte temporal de los datos para robustecer el entrenamiento.
- Implementar técnicas de selección de variables adicionales (LASSO, Sequential Feature Selection).
- Integrar los modelos a un sistema de actualización mensual, asegurando que reflejen cambios en los procesos productivos

Estas líneas de mejora son coherentes con las recomendaciones de la literatura. Rodríguez et al. (2022) sugieren que la incorporación de variables externas (mezcla operativa, tráfico o condiciones ambientales) aumenta la capacidad de generalización de los modelos logísticos, mientras que Singh (2021) destaca la importancia de ciclos de actualización continua para evitar degradación del desempeño por cambios en la operación.

La implementación progresiva de estas mejoras tendría un impacto directo tanto en la eficiencia operativa como en la confiabilidad del modelo predictivo. La incorporación de nuevas fuentes de información, el reentrenamiento periódico y la integración con

sistemas de captura en tiempo real permitirían reducir la incertidumbre del proceso, mejorar la capacidad de generalización del modelo y fortalecer la toma de decisiones basada en datos. En conjunto, estas acciones incrementarían la precisión de las estimaciones y consolidarían el valor práctico de la solución dentro de la operación logística de la planta.

CAPÍTULO 5

5. CONCLUSIONES

Las conclusiones del proyecto se estructuran en función del objetivo general y de los objetivos específicos planteados al inicio del estudio.

1. Cumplimiento del objetivo general.

El objetivo general de *implementar una herramienta de visualización interactiva basada en modelos de aprendizaje automático que permita apoyar la planificación logística y mejorar la eficiencia del proceso de carga* se cumplió de forma satisfactoria. Se desarrolló un modelo predictivo de tiempo de carga integrado en un dashboard de Power BI, que brinda al planificador una estimación más precisa de la duración de cada transporte y una visualización clara de las desviaciones. Esta herramienta convierte la información histórica en un insumo accionable para la toma de decisiones diarias.

2. Análisis de los datos operativos.

El análisis exploratorio permitió caracterizar la operación de la planta y comprender la alta variabilidad de los tiempos de carga. Se identificó la influencia del peso real, del flujo operativo y de la mezcla de presentaciones sobre la duración del proceso, así como la existencia de desviaciones mensuales significativas. Este entendimiento de los datos fue clave para seleccionar las variables más relevantes y sentar las bases del modelado, en línea con las recomendaciones de la literatura en proyectos de analítica logística.

3. Entrenamiento y validación de modelos predictivos.

Se evaluaron distintos algoritmos (Regresión Lineal, Random Forest y XGBoost), demostrando que Random Forest ofrece el mejor balance entre sesgo y varianza. El desempeño fue estable en distintos segmentos del conjunto de prueba, confirmando la fiabilidad del modelo en escenarios operativos con alta variabilidad.

4. Implementación en Power BI.

La integración del modelo predictivo en Power BI permitió desarrollar una herramienta visual interactiva para el planificador logístico. El dashboard facilita la comparación entre tiempos reales y estimados, permite identificar desviaciones y brinda soporte a la toma

de decisiones. Este resultado aporta valor inmediato a la operación y mejora la visibilidad del proceso.

5. Evaluación del impacto operativo y económico.

El modelo predictivo de Random Forest demostró una mejora sustancial en la precisión de la estimación de tiempos de carga, logrando una reducción del error promedio (MAE) de 47.3 % (pasando de 14.02 a 7.38 minutos). Este aumento en la fiabilidad es la base del impacto operativo y económico.

El modelo permitió reducir entre 40 y 70 horas operativas mensuales y disminuir costos directos entre USD 1,200 y USD 1,600 por mes. El impacto anual estimado es cercano a USD 20,040 sin considerar beneficios indirectos como el aumento del indicador OTIF, la reducción de tiempos muertos y la mejora en la estabilidad del proceso productivo.

6. Limitaciones y oportunidades de mejora.

El desempeño del modelo depende de la calidad y consistencia de los datos operativos. Factores externos no considerados —como condiciones climáticas, cambios en la mezcla de productos o daños por mantenimiento— pueden afectar la capacidad de generalización. Por ello, se recomienda un ciclo de actualización periódica del modelo y la incorporación futura de nuevas fuentes de datos.

7. Proyección estratégica.

Este proyecto constituye un primer paso hacia la implementación de un sistema de planificación inteligente basado en datos. La integración futura con una integración de SAP con una base de datos, servidores, reentrenamiento automático y alertas tempranas podría potenciar aún más la eficiencia logística y consolidar una cultura de toma de decisiones basada en evidencia.

Desde una perspectiva académica, este estudio aporta evidencia empírica sobre la aplicación de la metodología CRISP–DM en un entorno logístico real del sector agroindustrial ecuatoriano, mostrando cómo la integración entre modelos de aprendizaje automático y herramientas de visualización puede mejorar la toma de decisiones operativas. Esta contribución es relevante para futuros trabajos en analítica aplicada a operaciones, tanto en el ámbito local como regional.

Como recomendaciones prácticas, se sugiere institucionalizar el uso del dashboard en la planificación diaria, así como establecer un proceso mensual de actualización del modelo predictivo con nuevos datos operativos. A futuro, se propone explorar la

integración con sensores IoT, incorporar variables exógenas que actualmente no se registran y evaluar modelos secuenciales que consideren explícitamente la dinámica temporal del proceso de carga.

Finalmente, se destaca que todos los datos utilizados fueron tratados bajo criterios de confidencialidad, respetando la privacidad de la operación y utilizándose exclusivamente con fines de análisis y mejora interna. En conjunto, el proyecto demuestra que la integración de modelos de aprendizaje automático con herramientas de visualización interactiva puede transformar la planificación logística en procesos industriales reales. La reducción significativa del error, el ahorro operativo y la mejora en la visibilidad del proceso validan la utilidad práctica de la solución propuesta y consolidan su aportación metodológica al ámbito de la analítica aplicada.

BIBLIOGRAFÍA

1. Araujo, R., & Etemad, S. (2020). *Predictive analytics for logistics optimization in agro-industrial operations*. *Journal of Operations Engineering, 14*(2), 45–57. https://doi.org/10.1109/JOE.2020.114578
2. Ballou, R. H. (2004). *Business logistics/supply chain management: Planning, organizing, and controlling the supply chain* (5th ed.). Pearson Prentice Hall.
3. Banco Interamericano de Desarrollo. (2020). *Transformación digital y analítica de datos en la logística latinoamericana*. BID.
4. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Brzozowska, A., Kowalski, P., & Milani, F. (2023). Application of CRISP-DM for predictive modeling in small-batch manufacturing. *International Journal of Production Research, 61*(11), 3772–3791. https://doi.org/10.1080/00207543.2022.2145789
6. CEPAL. (2019). *Logística y competitividad en América Latina y el Caribe*. Comisión Económica para América Latina y el Caribe.
7. CEPAL. (2020). *Logística, transporte y transformación productiva en América Latina*. Comisión Económica para América Latina y el Caribe.
8. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc. https://www.the-modeling-agency.com/crisp-dm.pdf
9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
10. Christopher, M. (2016). *Logistics & supply chain management* (5th ed.). Pearson Education.
11. FAO. (2021). *Eficiencia operativa y logística en cadenas agroindustriales*. Organización de las Naciones Unidas para la Alimentación y la Agricultura.
12. Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten* (2nd ed.). Analytics Press.
13. Gattorna, J. (2019). *Dynamic supply chain alignment: A customer value-driven approach* (3rd ed.). Routledge.

14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
15. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
16. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.
17. INEC. (2022). *Indicadores de productividad del sector industrial ecuatoriano*. Instituto Nacional de Estadística y Censos.
18. Landín, M., & Reina, A. (2021). Implementación de CRISP-DM para mejorar la planificación operativa en centros logísticos. *Revista Iberoamericana de Ingeniería Industrial, 12*(3), 85–98. https://doi.org/10.33412/riei.2021.03.85
19. Microsoft. (2023). *Power BI documentation*. https://learn.microsoft.com/power-bi
20. Ministerio de Producción, Comercio Exterior, Inversiones y Pesca. (2021). *Agenda de productividad y competitividad del sector logístico en el Ecuador*. Gobierno del Ecuador.
21. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). *Introduction to linear regression analysis* (5th ed.). Wiley.
22. Plotnikova, V., Dumas, M., & Milani, F. (2022). Adapting CRISP-DM for privacy-preserving data mining in the banking sector. *Decision Support Systems, 163*, 113693. https://doi.org/10.1016/j.dss.2022.113693
23. Rodríguez, L., Martínez, V., & Ramírez, J. (2022). Predictive modeling for loading time estimation in fertilizer plants using decision trees. *Latin American Journal of Industrial Engineering, 9*(1), 44–56. https://doi.org/10.5958/LAJE.2022.9.44
24. Rushton, A., Croucher, P., & Baker, P. (2022). *The handbook of logistics and distribution management* (7th ed.). Kogan Page.
25. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
26. Singh, A. (2021). Machine learning models for truck arrival time prediction in distribution centers. *International Journal of Logistics Systems and Management, 38*(4), 512–530. https://doi.org/10.1504/IJLSM.2021.115784
27. Zhou, X., & Wang, Y. (2019). Delay prediction in Chinese logistics hubs using Random Forest models. *Transportation Research Part E: Logistics and Transportation Review, 129*, 134–149. https://doi.org/10.1016/j.tre.2019.06.001