PREDICCIÓN DE CAPACIDAD Y EFICIENCIA EN PLANTA DE PRODUCCIÓN DE ESPECIALIDADES QUÍMICAS MEDIANTE EL ANÁLISIS Y MODELADO AVANZADO DE DATOS.

Daniel Felipe Duarte Quintero.
Código 8986557

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Directora
Isabel Cristina García Arboleda

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JULIO 06 DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

En el sector de manufactura de especialidades químicas, la discrepancia entre la capacidad teórica y la capacidad real de producción representa una problemática crítica con implicaciones directas sobre la eficiencia operativa, la rentabilidad y el nivel de servicio. En particular, durante el periodo comprendido entre junio de 2023 y junio de 2024, se evidenció una brecha promedio del 20% entre lo planeado y lo ejecutado en planta, atribuible a factores como tiempos muertos no detectados, indisponibilidad técnica de los equipos y variabilidad en los procesos.

En el sector de manufactura de especialidades químicas, la brecha entre la capacidad teórica de producción y la capacidad real ejecutada constituye un desafío crítico con repercusiones directas sobre la eficiencia operativa, la rentabilidad y el cumplimiento de la demanda. En particular, Protecnica Ingeniería S.A.S. Una empresa colombiana con más de 47 años de experiencia en la fabricación de tensoactivos, emulsificantes, antiespumantes y otras especialidades químicas ha identificado una discrepancia promedio del 20% entre la capacidad planeada por tren y lo efectivamente producido durante el periodo junio 2023 - junio 2024.

Ante esta situación, el presente trabajo propone el diseño e implementación de un sistema predictivo integral que permite estimar, por tren de producción (agrupaciones tecnológicas definidas según la configuración de los equipos y la complejidad de los procesos), la capacidad mensual real alcanzable. Este sistema permite identificar posibles fallas de rendimiento y anticipar desvíos operativos. Se apoya en la recopilación automática de datos operativos, el análisis de variables clave, y la construcción de modelos de predicción utilizando técnicas de aprendizaje automático como Regresión Lineal Múltiple, Random Forest, XGBoost y Prophet.

A través de un pipeline desarrollado completamente en R, se consolida la información proveniente de múltiples fuentes (formularios operativos, plan de producción, datos históricos del ERP), se calculan indicadores como la disponibilidad (input del OEE), y se entrenan modelos por tren. Cada modelo es validado mediante métricas como RMSE, MAE, MAPE y R^2, tanto en conjunto de prueba como por validación cruzada (k-fold). Además, se construyó un tablero interactivo en Power BI que permite visualizar en tiempo real el estado operativo y las proyecciones por tren.

El objetivo general del proyecto fue predecir la capacidad real instalada mediante el modelado de datos operativos. Los objetivos específicos incluyeron: (i) consolidación y depuración de datos multi-fuente, (ii) desarrollo y comparación de modelos de predicción por tren, (iii) proyección a seis meses y comparación con la capacidad teórica, y (iv) visualización ejecutiva mediante herramientas de BI.

En suma, este enfoque representa un aporte significativo en el uso de ciencia de datos para la optimización de procesos industriales. Si bien la metodología fue aplicada a una planta específica de especialidades químicas, su arquitectura puede ser replicada o escalada a otros contextos productivos con condiciones similares.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En una empresa dedicada a la manufactura de especialidades químicas, se enfrenta una problemática estructural relacionada con la brecha persistente entre la capacidad teórica instalada y la capacidad real ejecutada. Esta discrepancia ha generado impactos negativos en la eficiencia operativa, provocando subutilización de recursos clave, cuellos de botella, incumplimientos de programación y pérdida de competitividad en el mercado.

Entre los factores que contribuyen a esta situación se encuentran la falta de trazabilidad precisa en tiempo real, el uso de sistemas manuales de captura de datos propensos a errores, y la ausencia de herramientas analíticas que permitan anticipar desviaciones operativas. Adicionalmente, la naturaleza por lotes de la producción y la diversidad tecnológica entre los trenes productivos acentúan la complejidad del entorno, limitando la posibilidad de utilizar enfoques estandarizados de planificación y control.

Desde una perspectiva de ciencia de datos, esta problemática plantea la necesidad de contar con un sistema predictivo que permita estimar, con mayor precisión, la capacidad alcanzable por tren de producción. Para ello, se diseñó una solución integral que incluye: la consolidación de datos históricos, la limpieza y estandarización de registros, la selección de variables operativas clave, el entrenamiento de modelos supervisados utilizando algoritmos de machine learning, y la validación rigurosa del desempeño predictivo. Este modelo se complementa con un tablero de control en Power BI, orientado a la visualización dinámica de indicadores y la generación de alertas tempranas.

Este enfoque no solo permite reducir la incertidumbre en la toma de decisiones estratégicas, sino también mejorar la asignación de recursos, identificar oportunidades de mejora táctica en la programación, y evaluar con mayor claridad el alineamiento entre la demanda proyectada y la capacidad efectiva de producción.

1.2 CONTEXTO ORGANIZACIONAL Y TRENES PRODUCTIVOS

El proyecto se desarrolló en una planta de manufactura de especialidades químicas que opera bajo un esquema de producción por lotes, con alta variabilidad en tiempos, materias primas, condiciones de operación y tecnologías involucradas. A diferencia de industrias con líneas de producción continua como envasado o ensamblaje automotriz donde es posible establecer una tasa constante de unidades por hora, en este tipo de operación cada lote responde a condiciones químicas y operativas particulares, lo cual dificulta la estandarización de tiempos y la modelación de capacidad de forma agregada.

Para gestionar esta complejidad, la planta ha estructurado su operación en trenes productivos, una clasificación funcional y tecnológica que permite agrupar los equipos según su infraestructura y el tipo de proceso químico que pueden ejecutar. Esta división responde a criterios de configuración técnica, compatibilidad energética y requerimientos del proceso, y ha sido clave para planear la producción, proyectar demanda por tipo de proceso y evaluar adecuadamente la capacidad instalada por tren, lo cual es fundamental para decisiones estratégicas, particularmente en el análisis de necesidades de inversión en infraestructura (CAPEX).

Cada tren está conformado por equipos con características específicas, que determinan qué tipo de procesos pueden realizar:

- Tren de esterificación: Procesos complejos que pueden durar entre 24 y 68 horas, que requieren calderas de aceite térmico para mantener temperaturas elevadas y estables. Los reactores asignados a este tren cuentan con chaquetas térmicas y conexiones específicas que permiten controlar de forma precisa las condiciones de reacción.
- Tren de reacción y mezcla en caliente: Requiere calderas de vapor y chaquetas de enfriaamiento, con tiempos operativos de 18 a 30 horas. Los reactores están conectados a sistemas de vapor a presión mediante válvulas, tuberías y controles adecuados para manejar temperaturas medias de forma eficiente.
- Tren de mezclas simples: No requiere calentamiento ni enfriamiento, y sus procesos suelen durar entre 4 y 12 horas. Los equipos de este tren están diseñados para realizar mezclas homogéneas a temperatura ambiente, sin infraestructura térmica asociada.
- Tren de mezclas en dos fases: Involucra la combinación de fases acuosas y oleosas procesadas por separado y unidas bajo condiciones específicas. Estos procesos, que toman entre 18 y 30 horas, requieren equipos con configuraciones auxiliares para manejo separado de fases y control de emulsificación.

Es importante enfatizar que los equipos no son intercambiables entre trenes. Cada reactor está físicamente conectado a un sistema energético y funcional particular (aceite térmico, vapor, torre de enfriamiento), y su diseño limita su uso a un único tipo de proceso. Por esta razón, la capacidad de cada tren está comprometida de forma estructural, y no se puede asumir que un aumento en la demanda de un tren pueda resolverse simplemente mediante reasignación de equipos.

En los últimos años, la empresa ha experimentado un crecimiento acelerado en su volumen de producción, impulsado por el aumento sostenido de la demanda y la diversificación de su portafolio de productos. Este dinamismo ha puesto en evidencia la necesidad de fortalecer la eficiencia operativa y la capacidad de anticipar posibles cuellos de botella en el sistema productivo. Aunque cada tren cuenta con una capacidad nominal teórica, en la práctica existen múltiples factores técnicos y logísticos que impiden alcanzar esos valores de forma sostenida. Por ello, es crucial contar con un modelo predictivo que permita proyectar de forma realista la capacidad mensual alcanzable por tren, compararla con la demanda proyectada y así tomar decisiones informadas sobre la carga, la planificación y las posibles inversiones requeridas. Este enfoque busca no solo optimizar el uso de los recursos existentes, sino también ofrecer a la organización una herramienta concreta para la toma de decisiones estratégicas, alineadas con los retos actuales de expansión y competitividad del negocio.

1.3 FORMULACIÓN DEL PROBLEMA

¿Cómo se pueden utilizar técnicas de ciencia de datos para predecir y optimizar la capacidad real de una planta de producción, mejorando la eficiencia operativa y apoyando la toma de decisiones estratégicas?

1.2.3 SISTEMATIZACIÓN

- ¿Cómo se debe obtener, depurar e integrar la información necesaria para alimentar un modelo predictivo de capacidad operativa?
- ¿Qué variables y algoritmos son adecuados para desarrollar un modelo que prediga la ocupación, disponibilidad y rendimiento operativo?
- ¿Cómo se puede evaluar y validar el modelo predictivo para asegurar su precisión y confiabilidad en el contexto de la planta?
- ¿Cómo se puede proyectar la capacidad futura con el modelo y cómo compararla con las metas teóricas definidas por la alta gerencia?
- ¿Qué elementos y estructura debe tener un tablero de mando para visualizar en tiempo real el desempeño y capacidad de los equipos?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Predecir la capacidad real instalada en la planta de producción de la compañía a través del análisis de datos históricos de ejecución de la producción, la identificación de las principales fallas, el seguimiento de los retrasos y la evaluación de la disponibilidad de los equipos para optimizar la eficiencia operativa y apoyar la toma de decisiones estratégicas informadas.

2.2 OBJETIVOS ESPECÍFICOS

- Recopilar y analizar información a través de la determinación del método de recolección de datos y el desarrollo de un modelo para su exploración y limpieza.
- Generar un modelo predictivo basado en datos históricos de disponibilidad, rendimiento y producción semanal para predecir la capacidad de la planta de producción.
- Evaluar y validar el modelo predictivo mediante técnicas de evaluación de modelos de aprendizaje automático o estadístico para garantizar su precisión y fiabilidad en la predicción de la capacidad de la planta de producción.
- Analizar las capacidades de las plantas de producción proyectadas con el modelo y compararlas con las capacidades teóricas que tiene la alta gerencia definidas.
- Desarrollar un tablero de mando a través de la integración de datos históricos y proyectados, basado en la recolección y análisis de información, la generación de modelos predictivos, y la comparación de capacidades, para monitorear en tiempo real el rendimiento, la ocupación y la disponibilidad de los equipos en las plantas de producción.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

La ciencia de datos se ha convertido en un eje fundamental para la transformación digital de las operaciones industriales. Su aplicación permite extraer conocimiento accionable a partir de grandes volúmenes de datos, mejorando procesos, optimizando recursos y generando ventajas competitivas [1]. En el contexto de plantas de producción, su potencial radica en la capacidad de predecir eventos, anticipar desviaciones y soportar la toma de decisiones estratégicas [2]. Aplicaciones exitosas han demostrado mejoras en eficiencia, mantenimiento predictivo, detección de fallas y planificación de producción.

La ciencia de datos abarca desde la captura de datos (sensores, formularios, ERP), su almacenamiento y procesamiento, hasta la aplicación de modelos matemáticos y estadísticos para generar valor. En este sentido, su uso en la industria química representa una oportunidad para convertir datos operativos en información táctica y estratégica. En particular, este enfoque se alinea con los principios de la Industria 4.0, donde la conectividad, el análisis de datos en tiempo real y la automatización inteligente se integran para optimizar la producción.

La ciencia de datos permite, además, establecer relaciones causales y correlacionales entre distintas variables operativas, optimizar recursos energéticos, y prevenir fallos críticos en el sistema productivo mediante técnicas de aprendizaje automático [3]. Las empresas que adoptan estas herramientas pueden pasar de un modelo reactivo a uno predictivo, logrando así ventajas competitivas sostenibles. En la literatura industrial, se han reportado casos de éxito donde la predicción de fallas, la optimización de rutas logísticas o la simulación de demanda han permitido reducir costos en hasta un 20% [4].

3.1.1 PLANIFICACIÓN DE CAPACIDAD EN ENTORNOS DE PRODUCCIÓN POR LOTES

La planificación de capacidad es una función clave en la gestión de operaciones, orientada a garantizar que los recursos de una planta sean suficientes para satisfacer la demanda proyectada de productos en un periodo determinado. Esta tarea se complejiza sustancialmente en entornos de producción por lotes, caracterizados por una alta variedad de productos, tiempos de procesamiento variables, y configuraciones tecnológicas específicas por tipo de proceso o equipo [5].

A diferencia de líneas de producción estandarizadas donde es posible establecer tasas de producción estables (por ejemplo, unidades por hora), en la producción por lotes cada orden de fabricación responde a especificaciones únicas: tipo de producto, fórmula, condiciones térmicas, duración del proceso, entre otras. Estas características hacen que la capacidad nominal de los

equipos no se traduzca automáticamente en capacidad real disponible, ya que múltiples factores operativos (cambios de línea, limpieza, indisponibilidad técnica, etc.) afectan el rendimiento efectivo [6].

En estos entornos, una práctica común es la segmentación de la planta por "trenes tecnológicos" o líneas funcionales, que agrupan equipos con capacidades similares y condiciones técnicas compatibles. Esta estrategia permite una mejor estimación del rendimiento real, dado que los equipos no siempre son intercambiables entre procesos, como ocurre cuando están configurados para operar con diferentes fuentes térmicas o líneas auxiliares específicas [7]. El modelado por trenes productivos, entonces, resulta útil para reflejar la estructura operativa real y evitar estimaciones agregadas que distorsionan la toma de decisiones.

Además, los enfoques clásicos de planificación como el MRP (Material Requirements Planning) o el RCCP (Rough-Cut Capacity Planning), si bien útiles en contextos repetitivos o con baja variabilidad, suelen ser insuficientes en plantas de especialidades químicas, donde la flexibilidad operativa y los tiempos de ciclo fluctuantes exigen herramientas más robustas. En este sentido, la aplicación de técnicas de machine learning para predecir capacidad real alcanzada permite incorporar datos históricos, condiciones de operación y comportamientos específicos por tren, mejorando la precisión en la estimación y aportando una base cuantitativa para la planeación operativa y estratégica [8].

Por tanto, en el presente trabajo se adopta un enfoque de modelado desagregado por tren productivo, considerando variables como horas operativas, disponibilidad, ocupación y complejidad del proceso, con el objetivo de construir un sistema predictivo que represente de forma más fiel la capacidad real mensual alcanzable por cada segmento operativo de la planta.

3.1.2 MACHINE LEARNING Y ALGORITMOS SUPERVISADOS / NO SUPERVISADOS

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial que permite a los sistemas aprender de los datos sin ser programados explícitamente. Existen dos grandes categorías: supervisado y no supervisado.

En el aprendizaje supervisado, el modelo aprende una función a partir de un conjunto de datos etiquetados, donde se conoce la variable objetivo (como la capacidad en kilogramos producidos). En el no supervisado, se busca encontrar patrones o estructuras ocultas en los datos sin una variable objetivo predefinida.

Entre los algoritmos supervisados más comunes están la regresión lineal, los árboles de decisión, Random Forest, XGBoost y las redes neuronales. Para tareas no supervisadas se encuentran el clustering (K-means, DBSCAN) y la reducción de dimensionalidad (PCA, t-SNE). Cada uno tiene sus fortalezas: por ejemplo, los modelos basados en árboles son robustos frente a datos ruidosos, mientras que las redes neuronales requieren grandes volúmenes de datos para ser efectivas.

El enfoque supervisado fue el eje de este proyecto, dado que se cuenta con registros históricos de variables y resultados. No obstante, también se usó clustering no supervisado (K-means) en etapas exploratorias [9]. Esto permitió identificar patrones ocultos y segmentar trenes productivos según su comportamiento operativo.

3.1.3 MODELOS SUPERVISADOS DE PREDICCIÓN

Se implementaron cuatro modelos de aprendizaje supervisado para estimar la capacidad real:

- Regresión Lineal Múltiple (RLM): modelo estadístico base ampliamente utilizado en ingeniería y econometría. Permite interpretar el impacto de cada variable independiente sobre la variable respuesta. Es fácil de explicar y calcular, aunque limitado ante relaciones no lineales o interacciones complejas [10]. Ha sido usado con éxito para predecir eficiencia energética, consumo de materias primas y tiempos de ciclo en procesos químicos.
- Random Forest (RF): técnica de ensamble que construye un conjunto de árboles de decisión sobre subconjuntos aleatorios del dataset. Destaca por su capacidad para manejar grandes volúmenes de datos, su resistencia al sobreajuste y su habilidad para estimar la importancia de las variables [11]. En entornos industriales es particularmente valioso cuando los datos contienen ruido o valores atípicos frecuentes.
- XGBoost: algoritmo de boosting basado en gradiente que construye árboles secuenciales corrigiendo errores de predicciones anteriores. Es uno de los algoritmos más precisos y eficientes para tareas de regresión, y se ha convertido en el estándar en competiciones de ciencia de datos [12]. Su capacidad para manejar interacciones complejas y su flexibilidad lo hacen ideal para sistemas industriales.
- Prophet: modelo desarrollado por Facebook orientado a la predicción de series de tiempo. Su estructura aditiva permite capturar estacionalidad, tendencias y eventos atípicos. Es especialmente útil para datos con estacionalidad no constante, como los de producción industrial [13]. A diferencia de ARIMA, Prophet es más amigable para usuarios sin formación estadística profunda.

Cada modelo aporta una perspectiva distinta: la RLM ofrece interpretabilidad; RF y XGBoost, alta capacidad predictiva en presencia de no linealidades; y Prophet, especialización en patrones temporales.

SUPUESTOS DE LA REGRESIÓN LINEAL MÚLTIPLE (RLM):

La regresión lineal múltiple se basa en una serie de supuestos fundamentales que deben ser validados para garantizar la confiabilidad del modelo. Entre ellos se encuentran: (i) linealidad entre las variables independientes y la respuesta, (ii) independencia de los residuos, (iii) homocedasticidad o varianza constante del error, (iv) normalidad de los residuos, y (v) ausencia

de multicolinealidad. La verificación de estos supuestos puede realizarse mediante pruebas estadísticas (como Shapiro-Wilk, Breusch-Pagan y VIF) y análisis gráficos (como Q-Q plots y residuos vs ajustados). El cumplimiento de estos principios asegura que las estimaciones y predicciones del modelo sean válidas y estables.

OPTIMIZACIÓN DE HIPERPARÁMETROS EN MODELOS DE APRENDIZAJE AUTOMÁTICO:

En algoritmos como Random Forest y XGBoost, el rendimiento del modelo puede depender significativamente de la configuración de hiperparámetros, como la profundidad máxima de los árboles (max_depth), el número de iteraciones (nrounds) o la tasa de aprendizaje (eta). La optimización de estos valores se realiza mediante procesos de búsqueda como grid search o random search, generalmente combinados con técnicas de validación cruzada. Esta etapa permite encontrar configuraciones que minimicen métricas de error (como el RMSE), mejorando el ajuste del modelo a los datos sin incurrir en sobreajuste.

Estas consideraciones metodológicas permiten evaluar no solo el rendimiento de los modelos en términos de error, sino también su validez estadística y estabilidad operativa. A continuación, se presentan los criterios utilizados para dicha evaluación.

3.1.4 EVALUACIÓN DE MODELOS PREDICTIVOS

Para comparar el desempeño de los modelos se utilizaron las siguientes métricas:

- RMSE (Root Mean Squared Error): penaliza errores grandes, apropiado cuando hay sensibilidad al riesgo o al desperdicio [14].
- MAE (Mean Absolute Error): proporciona una estimación directa y fácil de interpretar del error promedio absoluto.
- MAPE (Mean Absolute Percentage Error): permite expresar el error como porcentaje, facilitando comparaciones entre procesos con distintas escalas.
- R^2 (coeficiente de determinación): mide cuánta varianza de la variable objetivo está siendo explicada por el modelo.

Estas métricas se calcularon en los subconjuntos de entrenamiento y prueba, y en validación cruzada para garantizar robustez. La combinación de estas medidas permite evaluar precisión, estabilidad y generalización. Adicionalmente, es posible complementar con gráficas de residuos, matriz de correlación o curvas de aprendizaje.

3.1.5 VALIDACIÓN CRUZADA (CROSS-VALIDATION)

La validación cruzada k-fold permite dividir el dataset en k subconjuntos, entrenando el modelo en k-1 partes y validando en la restante, iterando hasta cubrir todos los bloques. En este proyecto se aplicó este método por tren, lo que permite observar la estabilidad del modelo en contextos particulares y con diferentes volúmenes de datos [15].

Su aplicación es esencial en entornos industriales donde los datos por línea o tren de producción son limitados, y donde es crucial validar que el modelo funcione en todos los contextos operativos posibles. La validación cruzada es una de las mejores prácticas para evitar sobreajuste, especialmente cuando se trabaja con series temporales o estructuras jerárquicas como trenes.

3.1.6 FEATURE ENGINEERING E IMPORTANCIA DE VARIABLES

La ingeniería de características permite aumentar el poder predictivo del modelo al transformar, combinar o seleccionar variables relevantes. En el proyecto se identificaron variables como disponibilidad, horas planificadas, eficiencia semanal y costos asociados.

La importancia de variables se midió usando Random Forest y XGBoost, observando qué predictores aportaban mayor reducción de error. Esto también sirvió como filtro para evitar sobreajuste y mejorar la interpretabilidad de los modelos. A nivel técnico, se pueden usar medidas como Gini, permutación de variables o SHAP (SHapley Additive exPlanations) para una interpretación más profunda del modelo.

3.1.7 CLUSTERING NO SUPERVISADO (K-MEANS)

Se utilizó clustering como parte del análisis exploratorio. K-means agrupó trenes productivos con comportamientos similares. Este paso fue clave para entender si había patrones comunes o diferenciadores entre trenes de alta y baja eficiencia. Aunque no alimentó directamente el modelo predictivo, permitió segmentar y visualizar mejor la base de datos [16]. Esta metodología también puede emplearse para identificar turnos con mayor variabilidad, operadores críticos o días con mayor incidencia de fallas.

3.1.8 BUSINESS INTELLIGENCE (BI)

Business Intelligence es un conjunto de procesos, herramientas y metodologías que permiten

transformar datos en información estratégica. En la industria, su aplicación permite el monitoreo constante de indicadores clave (KPIs), generación de alertas tempranas y optimización de recursos a través de dashboards interactivos [17].

Power BI fue la herramienta seleccionada en este proyecto por su versatilidad, integración con Google Sheets, SQL y APIs, y su facilidad de uso por personal de diferentes niveles de la organización. El dashboard construido permite ver tendencias históricas, comparaciones entre trenes, proyecciones mensuales y diagnóstico visual de desviaciones. Esto facilita la toma de decisiones basadas en datos, en tiempo real, integrando los resultados del modelo predictivo con la operación diaria de la planta.

BI no solo mejora la visualización de los datos sino también la cultura organizacional hacia el uso de datos en la toma de decisiones. Según estudios recientes, las empresas que adoptan BI junto con modelos predictivos mejoran en un 15 a 25% su eficiencia operativa [18].

3.2 ANTECEDENTES

El avance hacia la digitalización en la industria manufacturera ha generado un creciente interés en la mejora de la eficiencia operativa a través de la implementación de sistemas automatizados de recopilación y análisis de datos. Investigaciones como la de Mouhib et al. han destacado la importancia de esta transformación digital al desarrollar sistemas de cálculo de OEE en tiempo real. Este enfoque, respaldado por estudios previos, subraya cómo la automatización de la recopilación de datos y su análisis inmediato pueden mejorar significativamente la precisión y utilidad de las métricas de OEE [19].

Sin embargo, este proyecto va más allá al integrar métodos avanzados de modelado y predicción para estimar el OEE y también la capacidad de las plantas de producción de la industria de especialidades químicas. Esta ampliación del enfoque refleja una comprensión más profunda de las necesidades del sector manufacturero, donde la capacidad de anticipar las fluctuaciones en la producción es crucial para una gestión eficiente. La inclusión de modelos predictivos y métodos analíticos avanzados representa una evolución natural en el contexto de la digitalización, permitiendo a las empresas no solo monitorear y visualizar el rendimiento en tiempo real, sino también anticipar y planificar proactivamente las operaciones futuras.

Al integrar la capacidad predictiva en el sistema, este proyecto proporciona a las empresas una herramienta integral para la toma de decisiones informadas en la gestión de la producción. Este enfoque holístico aborda no solo la eficiencia operativa en tiempo real, sino también la capacidad de adaptarse y responder de manera proactiva a los cambios en la demanda y las condiciones del mercado. En última instancia, esta convergencia entre la digitalización, la predicción de capacidad y la visualización del rendimiento refleja una tendencia emergente hacia la adopción de soluciones más avanzadas y completas en el sector manufacturero.

En 2024, Z. Mouhib [19] aborda el problema del monitoreo y mejora del OEE (Overall Equipment Effectiveness) en el contexto de la digitalización en la industria manufacturera. Reconoce que la medición precisa del OEE es esencial para la toma de decisiones informadas en la gestión de la producción, y destaca cómo los avances en tecnologías digitales y análisis de datos han permitido nuevos enfoques en este aspecto. Se mencionan estudios que exploran el impacto de las tecnologías de la Industria 4.0, como la recopilación automatizada de datos, sistemas de información en tiempo real, inteligencia artificial y aprendizaje automático, en la mejora del OEE. Además, se resalta la importancia de la digitalización en el cálculo del OEE, permitiendo el monitoreo en tiempo real y una toma de decisiones basada en datos.

El aporte principal del artículo radica en la propuesta de un sistema digitalizado y en tiempo real para el cálculo del OEE. Este sistema busca automatizar la recopilación y análisis de datos, permitiendo una evaluación precisa y continua del rendimiento del equipo. Destaca la integración de tecnologías como IoT, inteligencia artificial y computación en la nube para mejorar la eficiencia operativa y la toma de decisiones. Asimismo, resalta la importancia de la calidad y fiabilidad de los datos para optimizar la eficiencia en la producción.

En comparación con este proyecto, se puede observar una similitud en el enfoque hacia la digitalización y automatización del proceso de cálculo del OEE. Ambos reconocen la importancia de la precisión de los datos y su impacto en la toma de decisiones en la gestión de la producción. Sin embargo, este proyecto se diferencia al incorporar modelos y métodos predictivos para estimar la capacidad de la planta de producción, lo cual proporciona una ventaja adicional en la optimización de la eficiencia y la planificación de la producción. Además, destaca la implementación de un dashboard en Power BI para la visualización de los datos, brindando una herramienta adicional para el análisis y la toma de decisiones que generan valor agregado a la organización.

En 2018, Y. Ramírez [20] aborda el problema de la medición de la eficiencia en las líneas de producción mediante la implementación de un sistema de Productividad y Mejoramiento OEE. Identifica las líneas críticas de proceso, capacidades nominales instaladas y genera un listado de paradas de línea. Utiliza una plantilla en Microsoft Excel para la recolección de datos, diligenciada por operarios en formatos impresos.

En el trabajo se destaca la importancia de implementar modelos de OEE para el monitoreo de indicadores de productividad. Sin embargo, su enfoque se centra en la recolección de datos a través de formatos impresos y su posterior registro en Excel.

En contraste, este proyecto mejora este proceso mediante la implementación de tabletas con formularios en Google Forms, permitiendo la recolección de datos en tiempo real y su posterior análisis y modelado predictivo. Es importante destacar que, a diferencia del estudio de Ramírez, aquí se consideran los datos históricos recolectados para predecir las capacidades de la planta de

producción. Además, se contempla la visualización de la información recolectada en plataformas de Business Intelligence como Power BI.

En 2021, M. Díaz [21] aborda la toma de decisiones informadas para mejorar la eficiencia operativa en una planta de producción de equipos de seguridad. Utiliza LabView para la recolección y adquisición de datos, Microsoft SQL para la base de datos y el proceso de ETL, y Microsoft Power BI para la visualización de informes en tiempo real.

El estudio se centra en la toma de decisiones informadas para mejorar la eficiencia operativa mediante procesos de ETL y la generación de informes de Business Intelligence.

Aunque comparte similitudes en el desarrollo de un dashboard en Power BI para monitorear el rendimiento y las principales paradas de planta, difiere en la recolección de datos y su enfoque en la toma de decisiones. Además, no considera los datos históricos recolectados para predecir las capacidades de la planta de producción, a diferencia de este proyecto.

Teniendo como base los fundamentos conceptuales presentados, en el siguiente capítulo se aborda el desarrollo del sistema predictivo, desde la captura y estructuración de datos hasta el análisis exploratorio, etapa fundamental para preparar la información que alimentará los modelos descritos.

4 PREPARACIÓN, ESTRUCTURA Y DISEÑO DEL SISTEMA PREDICTIVO

Este capítulo describe, paso a paso, el diseño, construcción y despliegue del sistema de ciencia de datos desarrollado para resolver el problema de predicción de capacidad productiva en una planta

de especialidades químicas, considerando las particularidades de sus trenes productivos y la alta variabilidad de sus procesos por lotes. A diferencia de proyectos basados en fuentes de datos preexistentes o estructuradas, este trabajo exigió desarrollar desde cero toda la arquitectura técnica del sistema: desde los protocolos de captura operativa, integración de información y validación de variables clave, hasta los componentes analíticos y de visualización necesarios para su explotación. Lo que sigue detalla esta construcción metodológica, orientada a brindar una herramienta real de soporte a la toma de decisiones operativas y estratégicas.

4.1 DIAGNÓSTICO DEL ENTORNO Y AUSENCIA DE DATOS

En la etapa inicial, se evidenció que la planta no contaba con sistemas de captura automatizada de datos operativos. Las hojas de control eran físicas, el registro de eventos como paradas y etapas era inconsistente, y no existía trazabilidad digital entre lotes, trenes y disponibilidad. Esta situación limitaba cualquier posibilidad de modelar capacidad productiva a mediano o largo plazo.

4.2 DISEÑO DEL SISTEMA DE RECOLECCIÓN

Para habilitar un flujo de datos confiable, se definió un esquema de captura digital basado en Google Forms. Se diseñaron formularios específicos por tipo de evento (producción, parada, ajustes), los cuales fueron desplegados en tabletas industriales resistentes al ambiente de planta, una por cada tren productivo.

- Se capacitaron más de 20 operarios en su uso.
- Se establecieron turnos responsables del diligenciamiento.
- Los formularios fueron validados con supervisores y jefes de producción.

El backend fue centralizado en Google Sheets, permitiendo integraciones automáticas vía R.

Figura 1 Formulario de Google Forms (Autoría propia)

Este formulario comprende varias secciones de recolección de datos claves para su posterior análisis como lo son:

Primera visual: Usuario que registra y a que planta de producción pertenece.
Segunda visual: Producto, lote, reactor, kilogramos, fecha y hora de la etapa.
Tercera visual: Etapa del proceso (cargue, reacción, ajuste, descargue).

El formulario esta diseñado para que los operarios al seleccionar la planta de producción solo le permiten visualizar y seleccionar los productos y los reactores que están asignados a esa planta, esto con el objetivo de minimizar los errores a la hora de registrar.

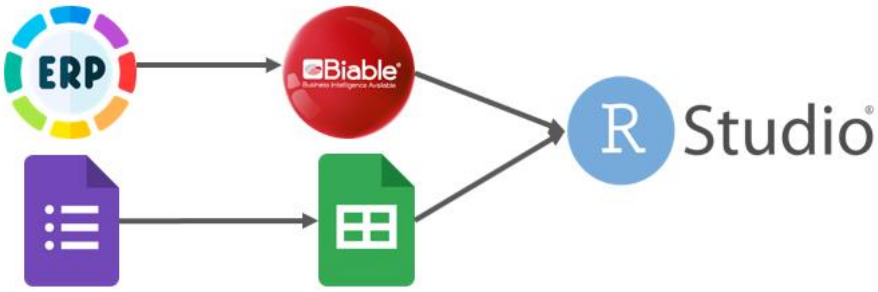

Figura 2 Flujo de los datos. Autoría propia

El flujo de datos de la recolección de información comprende desde la recolección de la de los históricos de producción del ERP y la información suministrada por los operarios a través del formulario de Google forms hasta la conexión mediante una API de Google Sheets a R Studio.

4.3 EXTRACCIÓN Y CONSOLIDACIÓN DE DATOS

Se construyó un pipeline en R con las librerías googlesheets4 y lubridate para extraer y estructurar más de 18.000 registros en tiempo real. Este proceso incluyó:

- Consolidación de múltiples formularios en una sola tabla.
- Identificación única de registros por lote, tren, fecha y etapa.
- Validación de duplicados y registros inválidos.
- Corrección de fechas mal registradas por los operarios.
- Se integró además información histórica desde el ERP de la compañía (datos de referencia como cantidad planeada por producto, códigos de ítem, etc.).

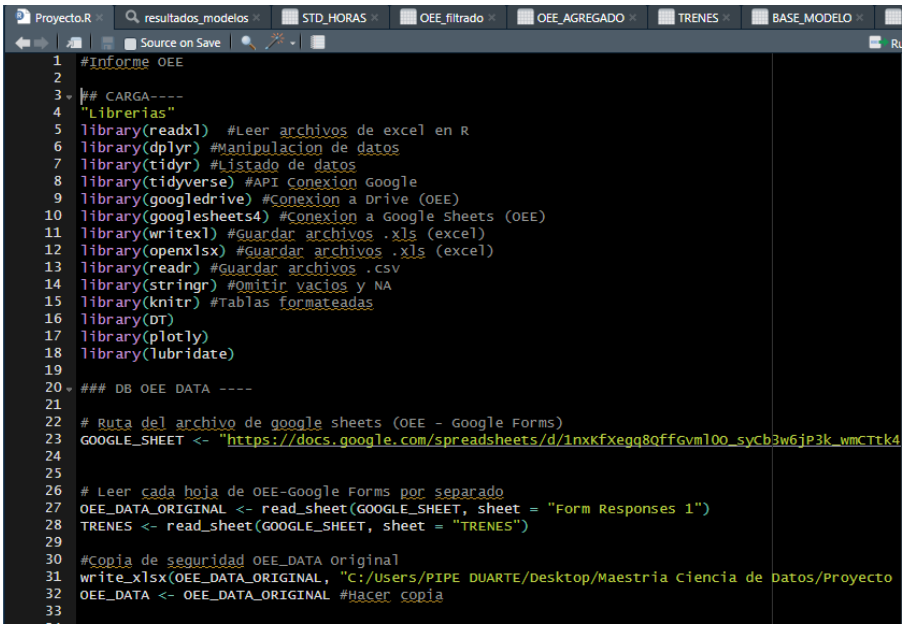

Figura 3 Conexión R Studio con Google Sheets (Autoría propia)

La figura 3 permite detallar las librerías que se emplearon en el proceso de recolección, extracción y procesado de la información en el software R Studio, durante esta etapa se generaron copias de seguridad de la base de datos original en Excel con el objetivo de garantizar la preservación de la información ante cualquier eventualidad.

```
> colnames(OEE_DATA)
 [1] "HORA REGISTRO"          "USUARIO REGISTRO"       "PLANTA DE PRODUCCIÓN" "PRODUCTO"
"REACTOR"                "LOTE"                   "CANTIDAD"
 [8] "HORA ETAPA"             "TIPO REPORTE"           "ETAPA"                "TIPO DE PARADA"
"CAUSA PARADA"           "TIPO DE PROCESO"        "REGISTRO COMPLETO"
[15] "PERIODO"               "SEMANA"                 "AÑO"                  "MES"
"TREN PRODUCTIVO"        "COMPAÑIA"               "ID_UNICO"
[22] "HORAS"
>
```

Figura 4 Data set OEE_DATA (Autoría propia)

Estas son las columnas del data set OEE_DATA sobre el cual se desarrolla este proyecto, esta información se obtiene a través de la librería googelsheets4 que permite la conexión en tiempo real entre R Studio y Google Sheets.

```r
## TRANSFORMACION ----
### DB OEE DATA ----


# Eliminar columnas
OEE_DATA <- OEE_DATA %>%
  select(-c(`PLANTA DE PRODUCCIÓN`, PRODUCTO_P1, PRODUCTO_P2, PRODUCTO_P3, PRODUCTO_P4, PRODUCTO_P5, PRODUCTO_P6,
            REACTOR_P1, REACTOR_P2, REACTOR_P3, REACTOR_P4, REACTOR_P5, REACTOR_P6,
            LOTE_BATCH, LOTE_FLOW,
            CANTIDAD_BATCH, CANTIDAD_FLOW,
            `HORA ETAPA_BATCH`, `HORA ETAPA_FLOW`,
            `TIPO REPORTE_BATCH`, `TIPO REPORTE_FLOW`,
            `ETAPA DEL PROCESO_BATCH [PROCESO]`, `ETAPA DEL PROCESO_FLOW [PROCESO]`, EQUIPO_ANT, LOTE_ETAPA, VALIDACION))

# Convertir columnas (formatos)
OEE_DATA$`USUARIO REGISTRO` <- as.character(OEE_DATA$`USUARIO REGISTRO`) # Convertir usuario registro a carácter (texto)
OEE_DATA$LOTE <- as.character(OEE_DATA$LOTE) # Convertir lote a carácter (texto)
OEE_DATA <- OEE_DATA %>% rename(`HORA REGISTRO` = Timestamp) # Renombrar columnas de Timestamp(hora registro sistema)
OEE_DATA <- OEE_DATA %>% rename(REACTOR = EQUIPO) # Renombrar columnas de EQUIPO a REACTOR
OEE_DATA <- OEE_DATA %>% rename(`PLANTA DE PRODUCCIÓN` = PLANTA) # Renombrar columnas de PLANTA A PLANTA DE PRODUCCIÓN
OEE_DATA <- OEE_DATA %>% rename(`TIPO DE PROCESO` = `TIPO PROCESO`) # Renombrar columnas de TIPO PROCESO A TIPO DE PROCE
OEE_DATA <- OEE_DATA %>%
  mutate(
    AÑO = substr(PERIODO, 1, 4),
    MES = substr(PERIODO, 5, 6)
  )

# Seleccionar y reordenar columnas
Orden <- c("HORA REGISTRO", "USUARIO REGISTRO", "PLANTA DE PRODUCCIÓN", "PRODUCTO",
           "REACTOR", "LOTE", "CANTIDAD", "HORA ETAPA", "TIPO REPORTE", "ETAPA",
           "TIPO DE PARADA", "CAUSA PARADA", "TIPO DE PROCESO", "REGISTRO COMPLETO", "PERIODO", "SEMANA", "AÑO", "MES")
OEE_DATA <- OEE_DATA %>% select(all_of(Orden))

# Reemplazar valores vacíos o NA en ETAPA, TIPO DE PARADA y CAUSA PARADA
OEE_DATA <- OEE_DATA %>%
  mutate(
    ETAPA = ifelse(is.na(ETAPA) | ETAPA == "", "PARADA", ETAPA),
    `TIPO DE PARADA` = replace_na(`TIPO DE PARADA`, "-"),
    `CAUSA PARADA` = replace_na(`CAUSA PARADA`, "-"),

    # Reemplazar "_" por "-" en LOTE
    LOTE = str_replace_all(LOTE, "_", "-"),

    # Eliminar "." y "," en LOTE
    LOTE = str_replace_all(LOTE, "[.,]", ""),

    # Convertir LOTE a minúsculas
    LOTE = str_to_lower(LOTE),
```

Figura 5 Código en R para transformación del data set (Autoría propia)

Este fragmento de código en R muestra la transformación del conjunto de datos capturado desde planta. Se eliminan columnas irrelevantes, se renombran variables para homogeneizar nombres, y se convierten formatos de texto y fecha. Además, se normalizan los campos de etapa, tipo de parada y causa, asegurando consistencia en los registros. Esta limpieza estructural se debe realizar para construir un data set confiable para modelado predictivo.

4.4 PREPARACIÓN Y TRANSFORMACIÓN DE DATOS

Como parte del análisis exploratorio, se incluyeron visualizaciones que permitieron caracterizar el comportamiento de los datos antes de aplicar los modelos. Las siguientes gráficas se usaron para evaluar la distribución y relaciones clave entre variables operativas:

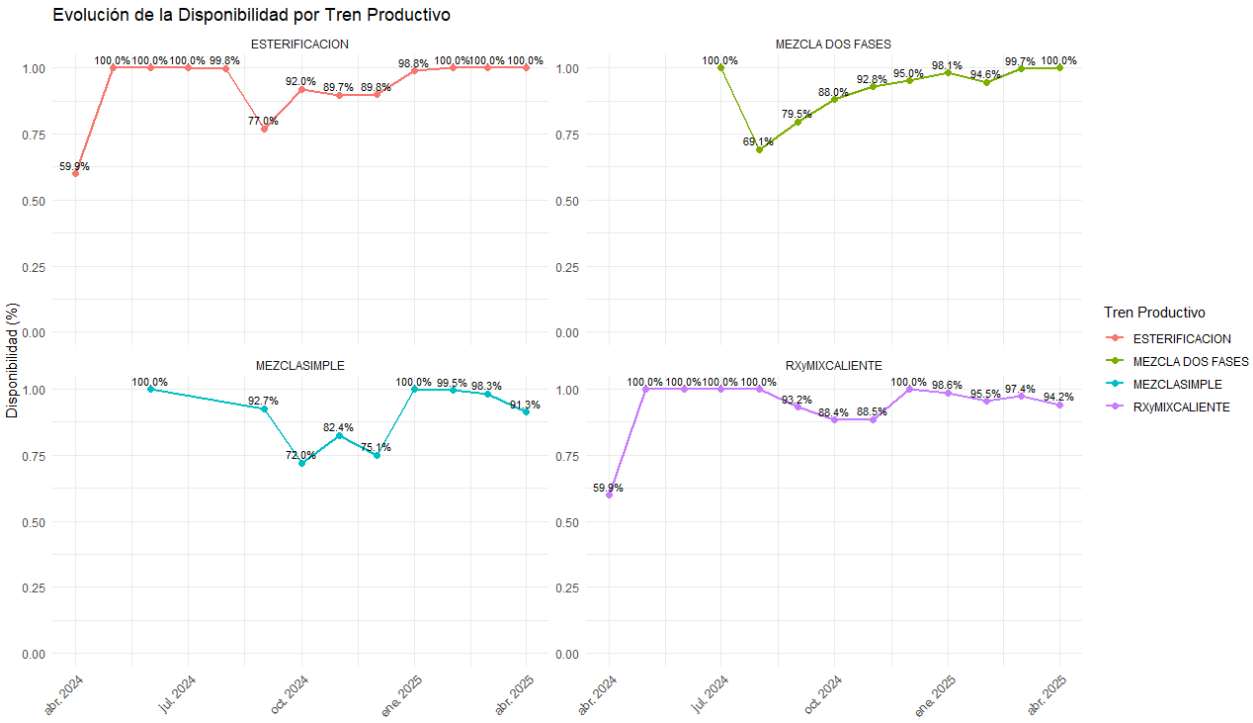

Figura 6 Grafico disponibilidad (Autoría propia)

Este gráfico muestra la evolución mensual del porcentaje de disponibilidad operativa por tren el cual es calculado mediante la fórmula de tiempo total de operación sobre el tiempo total programado. La disponibilidad es un componente clave del OEE y su análisis permite identificar trenes con mayores paradas no programadas o cuellos de botella.

Como se evidencia en las gráficas se presentan meses con ciertas fluctuaciones en los trenes (paradas no programadas) que afectan este indicador y este a su vez disminuye el output de la planta en ese tren determinado para ese periodo de tiempo.

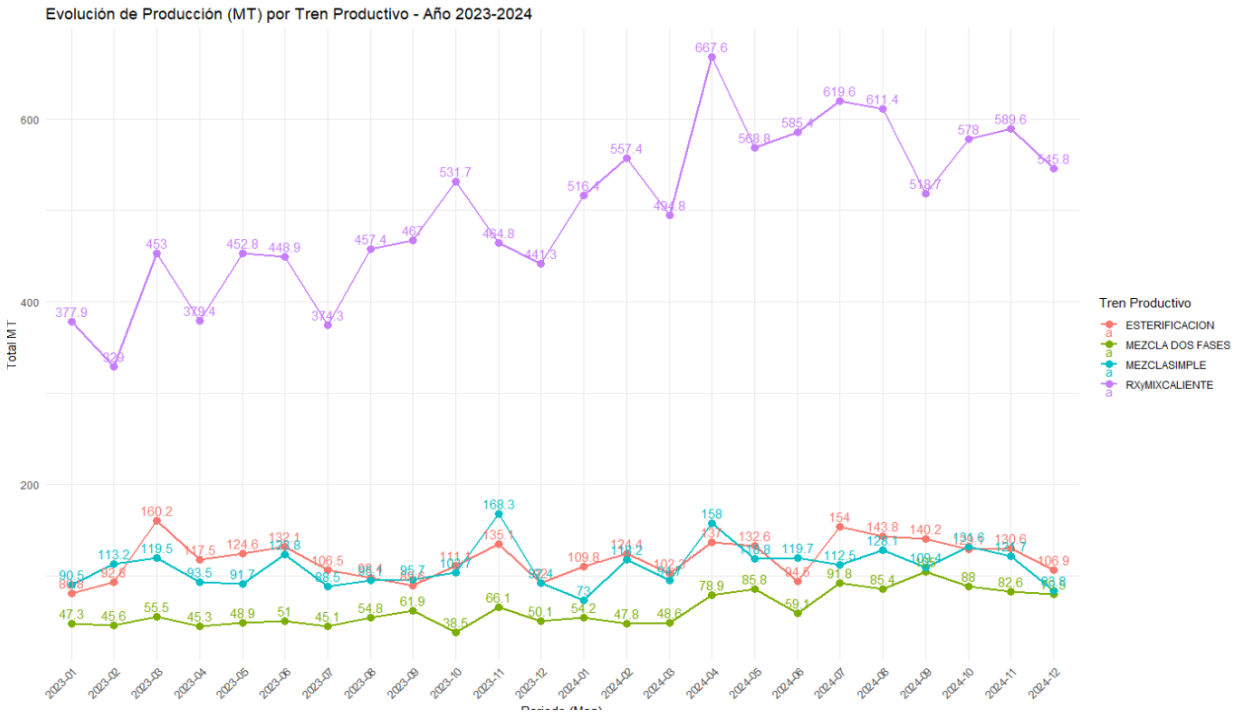

Figura 7 Gráfico histórico kilogramos fabricados por tren. (Autoría propia)

Se presentan las toneladas procesadas por cada tren productivo durante el período de estudio. Esto permite observar las diferencias de carga operativa entre trenes como reacciones y mezclas en caliente y mezclas de dos fases.

La figura presenta la producción total mensual en toneladas métricas por tren productivo. Se observa que el tren reacciones y mezclas en caliente concentra el mayor volumen con un comportamiento creciente hacia mediados de 2024, superando las 600 toneladas en múltiples meses. En contraste, los trenes de esterificación, mezcla dos fases y mezcla simple presentan volúmenes mucho menores y mayor variabilidad relativa. Esta diferencia justifica la necesidad de modelar individualmente cada tren, dada su heterogeneidad operativa.

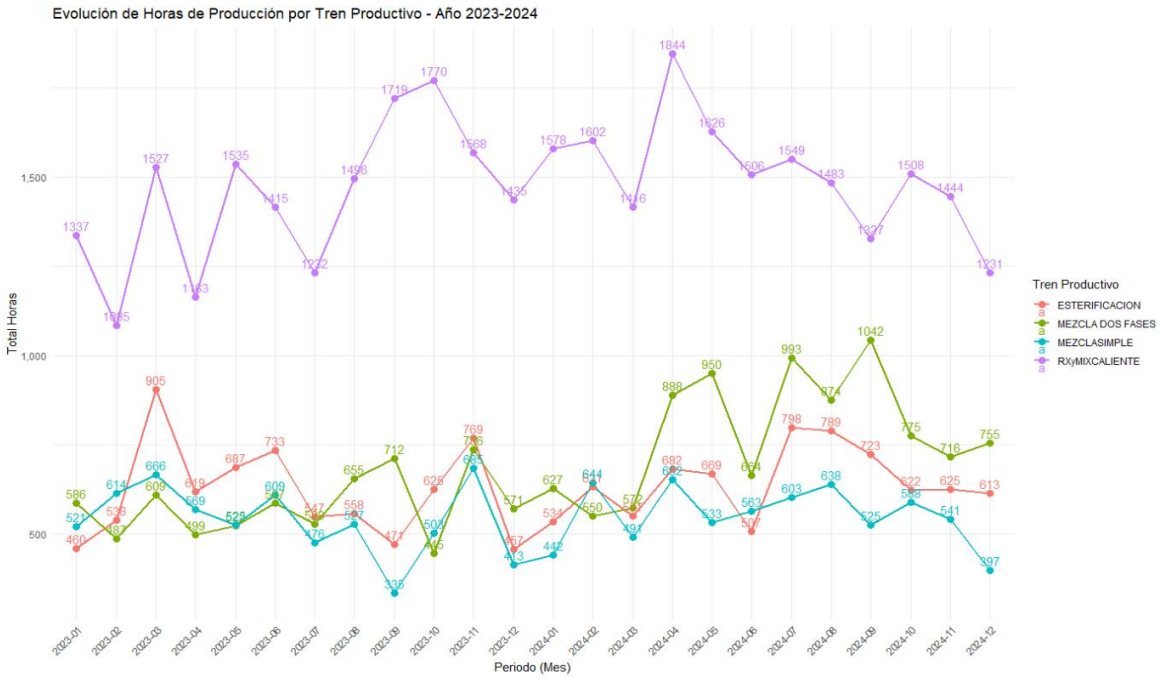

Figura 8 Gráfico histórico horas de fabricación por tren (Autoría propia)

La Gráfica permite visualizar el total de horas de producción por tren por mes y es fácilmente identificable que así como en los kilogramos fabricados donde más horas se utilizan los reactores es en reacciones y mezclas en caliente.

Esta gráfica muestra el total de horas de operación utilizadas mensualmente por cada tren. El tren reacciones y mezclas en caliente opera consistentemente con el mayor número de horas, alcanzando picos de hasta 1.844 horas en 2024. En contraste, esterificación y mezcla simple presentan menor intensidad horaria y mayor variabilidad, con caídas marcadas en varios periodos. Esta información permite comprender la carga operativa real y su impacto en la capacidad alcanzada.

Figura 9 Gráfica de correlación (kg vs horas) por tren (Autoría propia)

Esta Gráfica muestra la relación entre las horas de producción vs los kilogramos fabricados por tren, mostrando que trenes como esterificación y mezclas de dos fases requieren alta intensidad para fabricar menor cantidad de kilos contrario a lo que pasa con los trenes de reacciones y mezclas en caliente y mezclas simples que son trenes que cuentan con reactores con mayor capacidad nominal, tiempos de proceso mas cortos y mayor volumen requerido (demanda).

Estas visualizaciones facilitaron la comprensión del comportamiento histórico y validaron la necesidad de construir modelos diferenciados por tren productivo.

Tanto el análisis de correlación como la ingeniería de características formaron parte de la etapa exploratoria previa al modelado. El objetivo de estas actividades no fue predecir directamente, sino entender mejor las relaciones entre variables, detectar redundancias, e identificar qué atributos podrían tener mayor valor explicativo para los modelos posteriores.

La etapa de preparación incluyó también un análisis estadístico de correlación para entender las relaciones lineales entre las variables más representativas del conjunto de datos. Se construyó un mapa de calor que permitió identificar redundancias y colinealidad entre variables, de forma previa a la selección final para el modelado.

```r
### CORRELACION ----
"MAPA DE CALOR: CORRELACION DE VARIABLES"

library(ggcorrplot)

# Calcular matriz de correlación
cor_matrix <- cor(BASE_MODELO %>%
                    select(HORAS, KG, Costo_prom_ent, DISPONIBILIDAD),
                  use = "complete.obs")

# Crear el mapa de calor con mejor contraste
ggcorrplot(cor_matrix,
           method = "square",  # Celdas cuadradas en lugar de círculos
           type = "lower",  # Solo muestra la mitad inferior de la matriz
           lab = TRUE,  # Muestra los valores numéricos
           lab_size = 4,  # Tamaño de etiquetas
           colors = c("#6D9EC1", "white", "#E46726"),  # Azul-Negro-Rojo para negativo/neutro/positivo
           title = "Mapa de Calor de Correlación",
           outline.col = "gray")  # Contorno para mejor visibilidad
```

Figura 10 Codigo mapa de calor: correlación de variables (Autoría propia)

La ingeniería de características fue una etapa clave en la construcción del modelo predictivo. A partir de la base consolidada, se implementaron transformaciones y cálculos adicionales para derivar nuevas variables que capturaran con mayor fidelidad el comportamiento operativo de la planta. Entre las variables creadas se encuentran:

- Disponibilidad semanal por tren, calculada como la razón entre horas efectivas de producción y el total de horas calendario.
- Kilogramos producidos por mes y tren, como variable respuesta para estimar capacidad.
- Costo promedio por producto, a partir de históricos de materia prima, mano de obra y CIF.
- Horas planificadas y ejecutadas, como insumo para métricas de eficiencia y ocupación.

A continuación, se presenta el mapa de calor de correlación (Figura 11), que permitió identificar relaciones fuertes entre ciertas variables —como KG, horas y costos— y detectar otras con escasa correlación, como la disponibilidad, lo que sugiere la necesidad de modelos no lineales para su predicción.

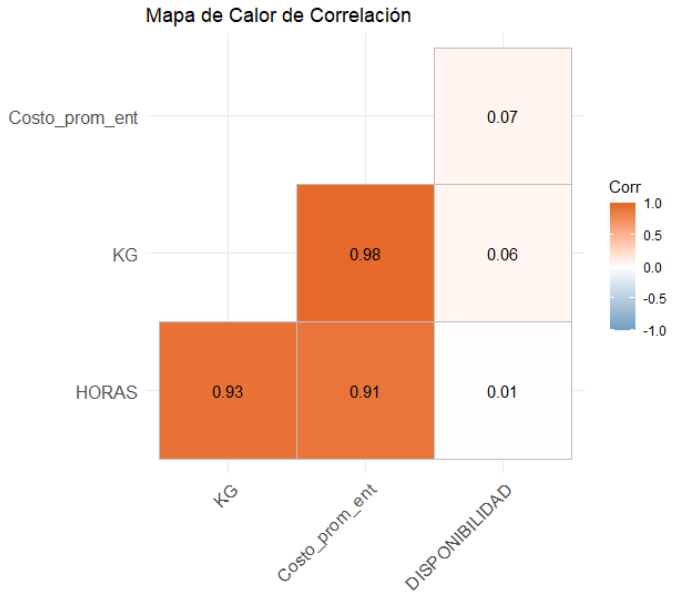

Figura 11 Mapa de calor: correlación de variables (Autoría propia)

Este mapa de calor revela alta correlación entre variables como KG, HORAS y COSTO_PROM_ENT, con coeficientes superiores a 0.90. En contraste, DISPONIBILIDAD muestra baja correlación lineal, lo que motiva el uso de modelos más flexibles como Random Forest o XGBoost para su predicción.

Para cuantificar la relevancia de cada variable en la predicción de la disponibilidad, se entrenó un modelo exploratorio con Random Forest. El script de implementación se muestra a continuación (Figura 12):

```r
### FEATURE ENGINEERING ----

"IMPORTANCIA DE LAS CARACTERISTICAS"
library(caret)
library(randomForest)
library(ggplot2)

# Eliminar las columnas 'Fecha' y 'TREN'
BASE_MODELO_SIN_FECHA_TREN <- BASE_MODELO[, !(colnames(BASE_MODELO) %in% c("Fecha", "TREN"))]

# Entrenar el modelo sin las variables 'Fecha' y 'TREN'
set.seed(123)
modelo_importancia <- train(DISPONIBILIDAD ~ ., data = BASE_MODELO_SIN_FECHA_TREN,
                            method = "rf", importance = TRUE)

# Obtener la importancia de las variables
importancia <- varImp(modelo_importancia)
df_importancia <- data.frame(Variable = rownames(importancia$importance),
                             Importancia = importancia$importance$Overall)

# Ordenar de mayor a menor
df_importancia <- df_importancia[order(df_importancia$Importancia, decreasing = TRUE), ]

# Graficar
ggplot(df_importancia, aes(x = reorder(Variable, Importancia), y = Importancia)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(Importancia, 1)), hjust = -0.1, size = 3.5) +
  coord_flip() +
  labs(title = "Importancia de Características (Random Forest)",
       x = "Variables", y = "Importancia") +
  theme_minimal()
```

Figura 12 Codigo Feature engineering (Autoría propia)

Los resultados del modelo exploratorio se visualizan en la Figura 13. Las variables con mayor importancia fueron:

- Costo_mo_en_ent (mano de obra),
- COSTO_MP_EN_ENT (materia prima),
- OCUPACION_KG,
- Costo_cif_en_ent y Costo_prom_ent.

Estas variables fueron priorizadas en la construcción de los modelos predictivos principales, mientras que otras con baja relevancia fueron descartadas para evitar ruido y mejorar la eficiencia computacional.

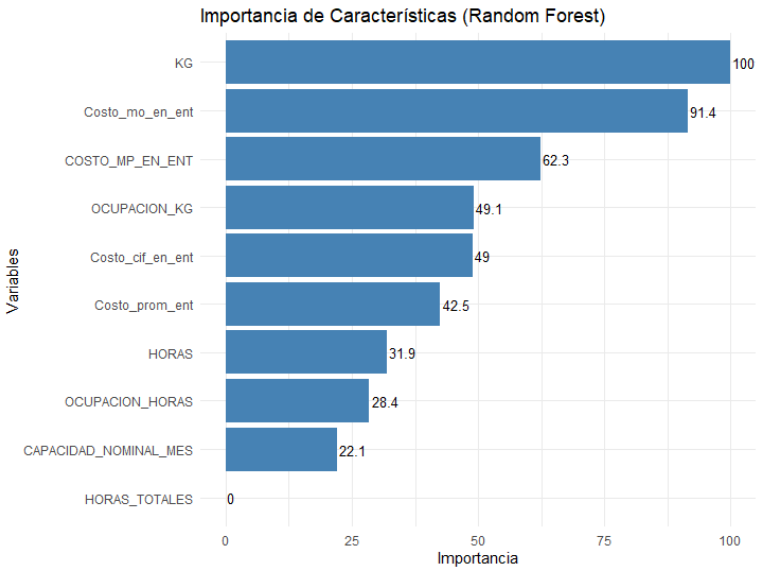

Figura 13 Gráfica Feature engineering (Autoría propia)

La gráfica presenta la importancia relativa de las variables predictoras en la estimación de la disponibilidad semanal, calculada mediante Random Forest. La variable KG (kilogramos producidos) resulta ser el predictor más influyente, seguido de Costo de mano de obra y Costo de materia prima en la entrada del inventario. También destacan métricas de ocupación y costos indirectos (CIF), lo que valida su inclusión prioritaria en los modelos principales. Variables como horas totales y capacidad nominal muestran menor contribución explicativa.

Una vez identificadas las variables más relevantes para predecir la disponibilidad mediante Random Forest, se procedió a calcular este indicador de manera sistemática a partir de los datos históricos de producción. Dado que la disponibilidad es uno de los componentes fundamentales del OEE, su estimación precisa por tren y periodo representa un insumo esencial para validar los modelos y entender las brechas operativas. A continuación, se presenta el script desarrollado para calcular este indicador a partir de los reportes de proceso y parada registrados por línea productiva.

```r
"Agrupar OEE_DATA para calculo de disponibilidad por TREN(linea)"
# Cargar las librerias necesarias
library(dplyr)
library(tidyr)
library(ggplot2)
library(lubridate)

# Filtrar los trenes específicos (ESTERIFICACION, MEZCLA DOS FASES, MEZCLASIMPLE, RXyMIXCALIENTE) y las columnas
trenes_interes <- c("ESTERIFICACION", "MEZCLA DOS FASES", "MEZCLASIMPLE", "RXyMIXCALIENTE")

OEE_filtrado <- OEE_DATA %>%
  filter(`TREN PRODUCTIVO` %in% trenes_interes, `TIPO REPORTE` %in% c("PROCESO", "PARADA"))  # Filtramos por los

# Crear una nueva columna de "Fecha" combinando MES y AÑO, y luego convertirla en un formato adecuado
OEE_filtrado <- OEE_filtrado %>%
  mutate(
    Fecha = as.Date(paste0(AÑO, "-", MES, "-01"))
  )

# Calcular las horas de proceso y horas de parada por mes, tren y año
OEE_AGREGADO <- OEE_filtrado %>%
  group_by(`TREN PRODUCTIVO`, Fecha, MES, AÑO, `TIPO REPORTE`) %>%
  summarise(
    HORAS = sum(HORAS, na.rm = TRUE),  # Sumar las horas por tipo de reporte
    .groups = "drop"
  ) %>%
  pivot_wider(names_from = `TIPO REPORTE`, values_from = HORAS, values_fill = list(HORAS = 0)) %>%
  mutate(
    HORAS_PROCESO = `PROCESO`,  # Asignar las horas de proceso
    HORAS_PARADA = `PARADA`,  # Asignar las horas de parada
    TOTAL_HORAS = HORAS_PROCESO + HORAS_PARADA,  # Calcular el total de horas
    DISPONIBILIDAD = HORAS_PROCESO / TOTAL_HORAS  # Calcular la disponibilidad (porcentaje de horas de proceso)
  ) %>%
  arrange(Fecha, `TREN PRODUCTIVO`)  # Ordenar por Fecha y Tren Productivo
```

Figura 14 Script en R para cálculo de disponibilidad (OEE) por tren (Autoría propia).

Este código filtra los registros por tren, tipo de reporte (“PROCESO” y “PARADA”) y estructura los datos por mes. Calcula las horas de proceso y de parada, y posteriormente estima el indicador de disponibilidad como el porcentaje de horas efectivas sobre el total disponible. Esta métrica constituye uno de los componentes principales del OEE, permitiendo evaluar la eficiencia operativa real por línea productiva. La disponibilidad es un factor que impacta la capacidad nominal instalada en la planta productiva por esta razón se debe calcular y evaluar.

4.5 ANÁLISIS EXPLORATORIO Y CLUSTERING

Como parte del análisis exploratorio previo al modelado, se implementaron técnicas de agrupamiento no supervisado con el algoritmo K-Means. El objetivo fue detectar patrones de comportamiento productivo a nivel de lotes y trenes, basados en variables operativas como producción (en toneladas métricas) y horas de operación.

El análisis permitió identificar distintos perfiles de comportamiento sin necesidad de etiquetas previas, facilitando una comprensión más profunda de la dinámica de planta. Si bien los resultados de clustering no fueron utilizados directamente como insumo en los modelos supervisados, sí ayudaron a justificar un enfoque desagregado por tren en lugar de uno global.

El número óptimo de clusters fue determinado mediante el método del codo, y se calculó un coeficiente de Silhouette promedio de 0.4658, lo que indica una estructura de agrupamiento moderadamente diferenciada, con cierta superposición entre grupos.

A continuación, se muestra el resultado gráfico del clustering:

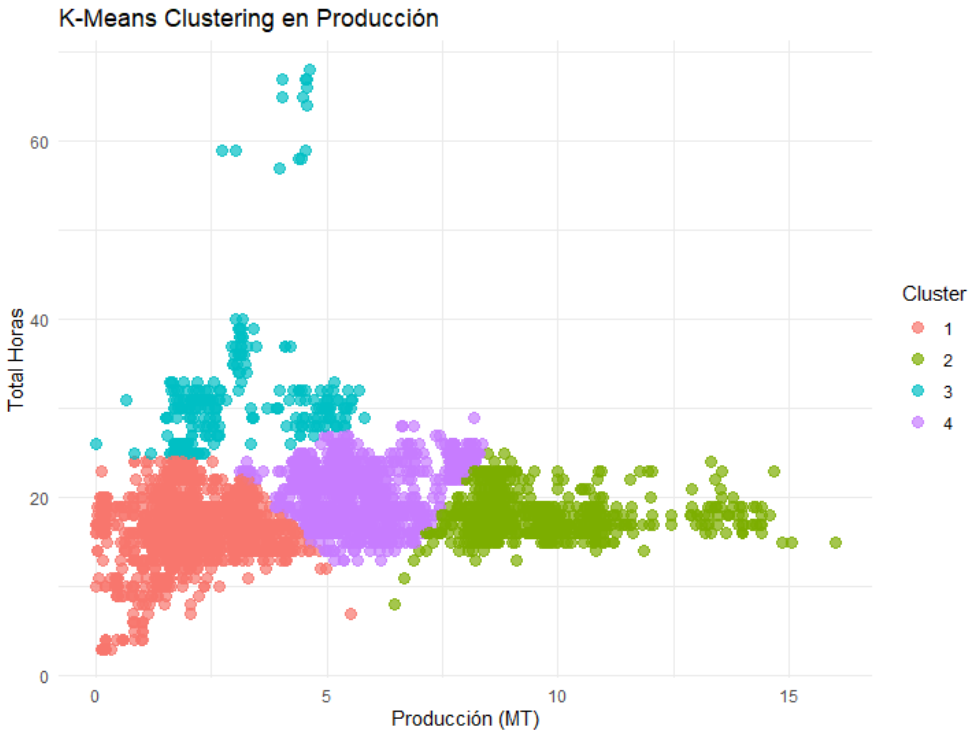

Figura 15 Gráfica K-Means (Autoría propia)

La figura muestra los cuatro grupos identificados según el comportamiento de los lotes. Cada punto representa un lote individual clasificado según su volumen de producción y sus horas totales asociadas. Se observan los siguientes patrones:

Cluster 1 (rojo): lotes con bajo volumen de producción y pocas horas requeridas.

Cluster 2 (verde): lotes eficientes, con alta producción y baja duración.

Cluster 3 (turquesa): lotes con baja producción y alta carga horaria (menos eficientes).

Cluster 4 (morado): lotes con producción media y duración moderada.

4.6 RESUMEN DEL PIPELINE DE INTEGRACIÓN Y LIMPIEZA DE DATOS

Como cierre del desarrollo de este objetivo, se presenta a continuación un resumen de las principales etapas del proceso de recolección, integración, limpieza y exploración de datos. Cada paso fue implementado mediante herramientas específicas que permitieron consolidar una base robusta para el posterior modelado predictivo.

ETAPA DEL PIPELINE	HERRAMIENTAS Y LIBRERIAS UTILIZADAS	PROPÓSITO
Recolección de datos operativos	Google Forms, Google Sheets	Capturar datos de producción en tiempo real desde planta
Integración y consolidación	googlesheets4, read_excel, left_join	Unificar registros de OEE, plan de producción y entregas
Limpieza y estandarización	stringr, lubridate, dplyr, replace_na	Homologar formatos, eliminar duplicados y corregir inconsistencias
Exploración y análisis preliminar	ggplot2, scales, ggcorrplot, facet_wrap	Visualizar tendencias y relaciones entre variables clave
Clustering exploratorio	kmeans, scale, cluster, silhouette	Identificar patrones comunes de comportamiento por lote o tren productivo
Ingeniería de características inicial	mutate, randomForest, caret, varImp	Calcular métricas derivadas y evaluar importancia de variables predictoras

Tabla 1 Resumen del pipeline (Autoría propia)

Una vez estructurado y validado el pipeline de captura, integración, limpieza y exploración de los datos operativos, se establece una base sólida para avanzar hacia el eje central de este proyecto: la construcción de modelos predictivos que estimen la capacidad real alcanzable por tren productivo. Esta transición marca el paso del análisis exploratorio descriptivo hacia la analítica predictiva, integrando el conocimiento obtenido en las fases previas para alimentar modelos robustos y contextualizados. En el siguiente capítulo se detallan las decisiones metodológicas, algoritmos utilizados y métricas aplicadas para evaluar su desempeño, con el fin de generar una herramienta confiable para la predicción y estimación de la capacidad instalada real de planta.

5 DESARROLLO DEL MODELO PREDICTIVO DE CAPACIDAD PRODUCTIVA

Con el objetivo de estimar la capacidad real máxima de planta en cada tren productivo, se desarrollaron modelos de predicción supervisados a partir de registros históricos operativos. Es importante destacar que la estimación de esta capacidad no corresponde a la producción efectivamente realizada, sino al potencial alcanzable bajo las condiciones operativas registradas, tales como la cantidad de horas disponibles, la disponibilidad de los equipos y ciertos factores económicos como los costos de materias primas y mano de obra. Esta distinción resulta clave, ya que el propósito es cerrar la brecha entre la capacidad teórica definida por la alta gerencia y aquella que realmente puede lograrse en planta considerando las condiciones actuales.

Se aborda con detalle el proceso técnico y analítico seguido para construir y comparar modelos por tren productivo. Se describen las etapas de preparación de datos, entrenamiento de modelos, validación cruzada, análisis de resultados y visualización analítica, con énfasis en identificar los factores que limitan o potencian la capacidad instalada real de planta.

5.1 PREPARACION DE LOS DATOS

La construcción del modelo predictivo se basó en una base consolidada, denominada BASE_MODELO, con datos mensuales por tren entre enero de 2023 y febrero de 2025. Esta base fue integrada a partir de diversas fuentes como los registros de producción efectiva (entregas de producción), la disponibilidad estimada a partir de reportes del OEE, los costos operativos reportados en el ERP, y las capacidades teóricas de cada tren. En esta base, la variable objetivo fue definida como los kilogramos fabricados (KG), utilizados como proxy para representar la capacidad alcanzada bajo un determinado conjunto de condiciones. Como variables explicativas o predictoras se utilizaron múltiples campos, entre ellos: disponibilidad mensual del tren (DISPONIBILIDAD), horas de operación efectivas (HORAS), costos promedio de producción (Costo_prom_ent), ocupación relativa respecto a la capacidad teórica (OCUPACION_KG y OCUPACION_HORAS).

Durante la etapa de preparación, se procedió a la imputación de datos faltantes en los primeros meses donde aún no se contaba con datos de OEE, utilizando el promedio observado en los meses más recientes con registros válidos. También se limpiaron registros atípicos, se corrigieron errores de codificación de trenes, y se realizó la estandarización de unidades para asegurar la homogeneidad del dataset. Finalmente, se dividió la base en un conjunto de entrenamiento (80%) y uno de prueba (20%) para evaluar la capacidad de generalización de los modelos. Esta fase concluyó con la generación de una matriz de correlación para identificar relaciones significativas entre variables, y un análisis exploratorio de importancia de características utilizando Random Forest.

5.2 MODELOS IMPLEMENTADOS

Para estimar la capacidad real alcanzable de cada tren productivo, se evaluaron varios algoritmos supervisados de predicción, priorizando aquellos que ofrecieran un balance entre precisión, interpretabilidad y adaptabilidad a las condiciones operativas de planta. Los siguientes modelos fueron seleccionados tras una fase exploratoria y de prueba de métricas:

Regresión Lineal Múltiple (RLM): se utilizó como modelo base de referencia, dada su simplicidad y facilidad de interpretación. Permitió identificar el peso de cada variable predictora (como disponibilidad, ocupación y costos) en la estimación de los kilogramos producidos. Aunque no fue el modelo con mejor ajuste, sirvió como punto de comparación fundamental.

Para garantizar la validez estadística del modelo de Regresión Lineal Múltiple (RLM), se evaluaron los supuestos clásicos: normalidad de residuos, homocedasticidad, colinealidad entre predictores y calidad de ajuste. Para cada tren productivo, se entrenó un modelo individual.

El resumen del diagnóstico estadístico de los modelos de regresión lineal múltiple se presenta en la siguiente tabla, discriminado por tren productivo. En ella se incluyen el coeficiente de determinación (R^2), los valores p de las pruebas de normalidad (Shapiro-Wilk) y homocedasticidad (Breusch-Pagan), así como el VIF promedio como indicador de colinealidad entre predictores. Estos resultados respaldan la validez estadística de los modelos utilizados y permiten verificar el cumplimiento de los supuestos fundamentales requeridos para su interpretación.

Diagnóstico de supuestos estadísticos por tren (Regresión Lineal Múltiple)

Tren	R^2	p (Shapiro-Wilk)	p (Breusch-Pagan)	VIF Promedio
ESTERIFICACION	0.8884	0.2459	0.5025	2.6514
MEZCLA DOS FASES	0.9480	0.7030	0.6787	1.9014
MEZCLASIMPLE	0.6192	0.6355	0.4203	2.1108
RXyMIXCALIENTE	0.7219	0.3740	0.6166	1.1887

Tabla 2 Diagnóstico de supuestos estadísticos por tren (Regresión Lineal Múltiple) (Autoría propia)

En general, los modelos mostraron un buen nivel de ajuste, con coeficientes R^2 superiores al 0.72 en casi todos los trenes, destacándose Mezcla dos fases con un R^2 de 0.948. La normalidad de los residuos fue evaluada con la prueba de Shapiro-Wilk, que en todos los casos arrojó p-valores superiores a 0.24, lo que indica que no se rechaza la hipótesis de normalidad. La homocedasticidad fue validada mediante la prueba de Breusch-Pagan, cuyos p-valores fueron superiores a 0.40 en todos los trenes, sugiriendo varianza constante de los residuos.

Respecto a la colinealidad entre predictores, se calculó el VIF promedio por modelo. En todos los casos se obtuvieron valores inferiores a 3, lo cual es aceptable según la literatura. El tren de esterificación presentó el mayor VIF promedio (2.65), aunque dentro de los rangos tolerables. Estos resultados respaldan la validez estadística general de los modelos RLM empleados, aunque se recomienda complementar con modelos no lineales en trenes con menor R^2 (por ejemplo, Mezcla simple con $R^2 = 0.62$).

Random Forest (RF): se aplicó por su capacidad para manejar relaciones no lineales, múltiples variables y datos con ruido, condiciones frecuentes en entornos industriales. Los modelos se entrenaron de forma independiente por tren, generando predicciones y rankings de importancia de variables. Se observaron buenos niveles de ajuste en la mayoría de trenes.

XGBoost: fue uno de los modelos con mejor desempeño, especialmente en trenes con mayor volumen de datos y menor variabilidad operativa. Su capacidad de corrección secuencial de errores y su eficiencia computacional lo hicieron especialmente útil para capturar relaciones complejas entre variables.

Prophet: se implementó por su capacidad para modelar series temporales con estacionalidad, tendencia y eventos atípicos, características comunes en la producción mensual por tren. Permitió generar proyecciones multimensuales y capturar ciclos de comportamiento recurrente. Fue particularmente útil en trenes con patrones estables.

Durante la fase exploratoria también se implementaron modelos clásicos de series de tiempo, incluyendo Promedio Móvil Simple (SMA), Holt, Holt-Winters y ARIMA, sin embargo, estos enfoques presentaron limitaciones importantes en su capacidad de ajuste: el coeficiente de determinación (R^2) fue inferior al 0.50 en la mayoría de los trenes, y las métricas de error como RMSE y MAE resultaron significativamente más altas que en los modelos seleccionados. Por esta razón, se decidió no incluir estos enfoques en el desarrollo final del modelo.

La elección de algoritmos se orientó así por su desempeño real sobre los datos históricos de planta, la facilidad de interpretación para usuarios operativos, y la posibilidad de generalizar sus resultados a periodos futuros bajo condiciones operativas similares.

Una vez seleccionados los cuatro modelos con mejor desempeño sobre los datos históricos (RLM, Random Forest, XGBoost y Prophet), se procedió a entrenarlos individualmente por tren productivo, con el fin de capturar las particularidades operativas de cada línea. Los modelos descartados en etapas previas —como SMA, Holt, Holt-Winters y ARIMA— fueron excluidos de

esta fase al mostrar bajos niveles de ajuste y métricas insatisfactorias durante la validación cruzada.

En esta etapa, cada modelo fue entrenado utilizando los registros mensuales de capacidad alcanzada y variables explicativas como horas operativas, ocupación, disponibilidad y costos. A partir de ello, se generaron predicciones de capacidad máxima alcanzable y se contrastaron con la capacidad teórica definida para cada tren. Esta comparación permite no solo evaluar el rendimiento de cada modelo, sino también identificar brechas entre el potencial proyectado y el límite nominal definido por la organización.

5.3 OPTIMIZACIÓN DE HIPERPARÁMETROS Y SELECCIÓN DE CONFIGURACIÓN FINAL

Para los modelos XGBoost y Random Forest, se implementó un proceso de ajuste de hiperparámetros mediante búsqueda por grilla (grid search) utilizando la función caret::train() con validación cruzada de tipo k-fold (k = 5). En el caso de XGBoost, la combinación óptima (bestTune) encontrada fue: nrounds = 50, max_depth = 2, eta = 0.1. Para Random Forest, el valor óptimo fue mtry = 4, dentro de un rango evaluado entre 2 y 5.

La métrica usada para seleccionar estos hiperparámetros fue el RMSE promedio en validación cruzada, y se fijó una semilla común (set.seed(123)) para garantizar la reproducibilidad de los resultados. Aunque los valores óptimos fueron registrados dentro de los objetos de entrenamiento (bestTune), se decidió no aplicarlos en el pipeline final por razones de robustez, simplicidad y consistencia operativa. Esta decisión se fundamenta en que, si bien en algunos trenes el ajuste mejoró ligeramente las métricas, en otros los resultados empeoraron, y el beneficio agregado no justificó la complejidad adicional. Esta evaluación detallada se discute en la sección 5.5.

5.4 RESULTADOS DE LOS MODELOS

A continuación, se presentan los resultados gráficos por modelo, en los que se visualiza la serie histórica de producción (línea azul), la proyección máxima estimada por el modelo (línea roja punteada) y la capacidad teórica de cada tren (línea verde punteada). Esto facilita una lectura integral del ajuste y utilidad de cada algoritmo en distintos contextos operativos.

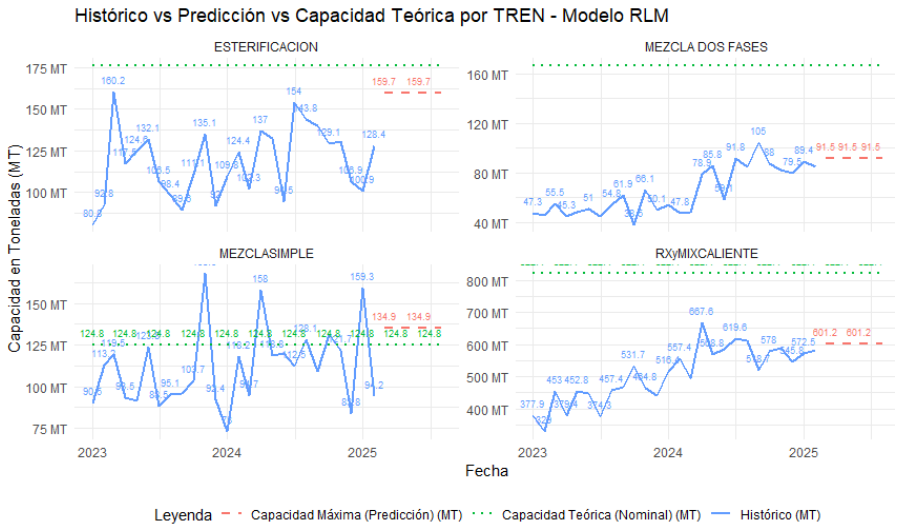

Figura 16 Proyeccion modelos RLM por tren (Autoría propia)

Esta figura muestra la proyección de capacidad máxima alcanzable por tren utilizando regresión lineal múltiple (RLM). Si bien el modelo logra capturar ciertas tendencias históricas, se observa una subestimación en trenes como reacciones y mezclas en caliente y mezcla de dos fases, donde la predicción máxima queda por debajo de la capacidad teórica. Esto evidencia que la RLM, aunque interpretable, puede ser limitada en entornos con relaciones no lineales o alta variabilidad.

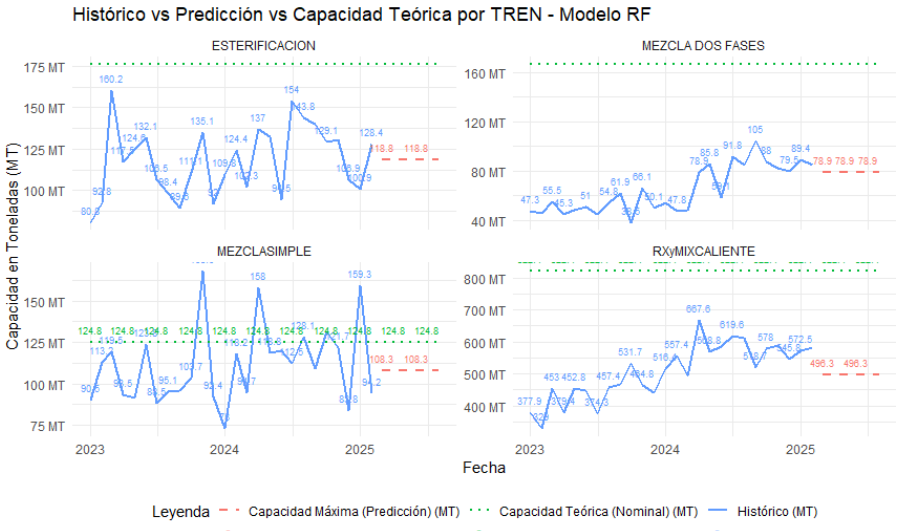

Figura 17 Proyeccion modelos RANDOM FOREST por tren (Autoría propia)

El modelo Random Forest mejora notablemente el ajuste respecto a RLM, especialmente en trenes como mezcla simple y esterificación. Sin embargo, sigue mostrando cierta subestimación en reacciones y mezclas en caliente. Las proyecciones se mantienen dentro de márgenes realistas y reflejan mejor la variabilidad mensual del histórico. Este modelo demostró robustez frente a outliers y buena capacidad de generalización.

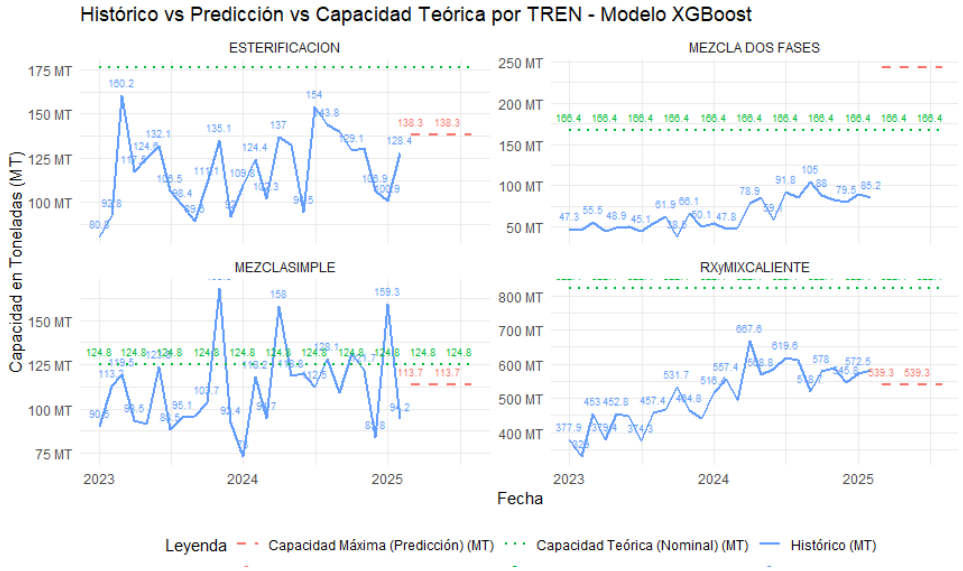

Figura 18 Proyeccion modelos XGBOOST por tren (Autoría propia)

XGBoost logra el mejor ajuste general, acercándose a la capacidad teórica en mezcla dos fases y capturando los picos de reacciones y mezclas en caliente con mayor precisión. Se destaca su capacidad para reflejar fluctuaciones rápidas y adaptarse a trenes con patrones menos lineales. Este modelo predice con mayor cercanía al histórico sin sobreajustar, lo que lo convierte en uno de los candidatos más sólidos para estimación futura.

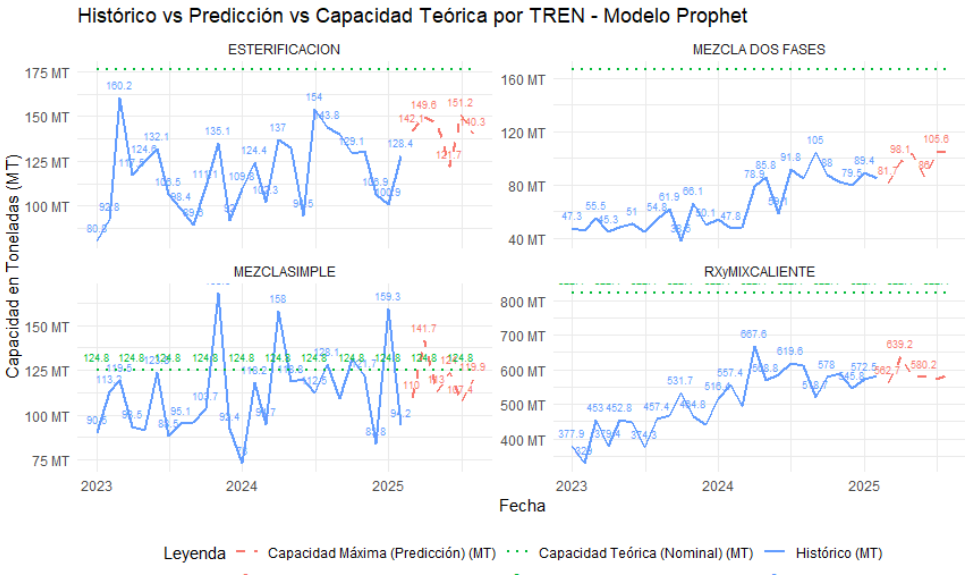

Figura 19 Proyeccion modelos PROPHET por tren (Autoría propia)

Prophet permite observar proyecciones multimensuales sobre la base de series temporales. Si bien su ajuste es aceptable en esterificación y mezcla simple, en trenes como reacciones y mezclas en caliente tiende a sobreestimar la capacidad alcanzable. Sus predicciones pueden verse afectadas por los cambios abruptos en comportamiento, lo que lo vuelve útil para líneas estables pero limitado para trenes con alta variabilidad o eventos atípicos.

Tras visualizar las predicciones por tren para cada uno de los modelos implementados, fue necesario complementar este análisis con una evaluación cuantitativa de desempeño. Para ello, se aplicaron métricas estándar que permiten comparar objetivamente la precisión y capacidad explicativa de cada algoritmo, tanto de manera global como específica por tren.

Esta etapa no solo permitió contrastar el ajuste de las proyecciones frente al histórico real, sino también identificar en qué contextos cada modelo mostró mejor comportamiento. Dado que el entorno operativo de cada tren presenta particularidades distintas como volumen, estacionalidad y variabilidad en horas, una evaluación desagregada resulta fundamental para seleccionar el modelo más adecuado en cada caso.

A continuación, se presentan los resultados de esta evaluación mediante métricas como RMSE, MAE, MAPE y R^2, aplicadas sobre los conjuntos de entrenamiento y prueba para cada modelo. Además, se incluye un análisis comparativo por tren para observar el comportamiento diferencial de los algoritmos según el perfil operativo de cada línea.

5.5 EVALUACION DE DESEMPEÑO

La evaluación se realizó mediante métricas estándar: RMSE, MAE, MAPE y R^2. Además, se aplicó validación cruzada tipo K-Fold para robustecer los resultados.

Estas métricas fueron seleccionadas por su capacidad para evaluar diferentes aspectos del error en problemas de regresión, donde la variable objetivo es continua (kilogramos de capacidad mensual alcanzada por tren productivo en kg).

- R^2 mide el poder explicativo del modelo sobre la variabilidad de los datos.
- RMSE penaliza errores grandes y es útil en contextos industriales donde desviaciones elevadas tienen alto impacto.
- MAE entrega un error promedio directo y robusto.
- MAPE permite expresar el error en términos porcentuales, lo que facilita su interpretación en trenes con diferentes escalas de producción.

No se utilizaron métricas como precisión, recall, F1-score o accuracy, ya que estas se aplican exclusivamente a tareas de clasificación, donde la variable objetivo es categórica. En este proyecto, el objetivo fue estimar una variable numérica continua (capacidad mensual alcanzada por tren), por lo que dichas métricas no resultan aplicables ni relevantes en el contexto de modelado de regresión.

Cabe resaltar que, aunque algoritmos como Random Forest y XGBoost pueden ser utilizados tanto en clasificación como en regresión, en este trabajo fueron implementados específicamente como modelos de regresión, ajustando una variable cuantitativa expresada en kilogramos. Por tanto, la evaluación de desempeño se basó en métricas adecuadas para este tipo de problema, como R^2, RMSE, MAE y MAPE, que permiten medir el poder explicativo y la precisión del modelo en un contexto numérico y continuo.

Para fortalecer la robustez de los modelos predictivos desarrollados, se aplicó validación cruzada tipo K-Fold, dividiendo la base de datos en subconjuntos de entrenamiento y prueba de manera sistemática. Esta técnica permitió evaluar la estabilidad de los modelos en diferentes particiones de los datos, minimizando el riesgo de sobreajuste y asegurando su capacidad de generalización. Su aplicación resulta especialmente crítica en entornos industriales tipo batch, donde la disponibilidad de datos históricos puede ser limitada o presentar alta variabilidad, garantizando así resultados más confiables y extrapolables a futuras condiciones operativas.

Adicionalmente, se realizó un análisis de desempeño global de los modelos considerando todos los trenes de producción de manera agregada, sin distinguir entre ellos. En esta evaluación global, los modelos mostraron coeficientes de determinación (R^2) superiores a 0.90 en su mayoría, reflejando un ajuste aceptable en términos de predicción general. Sin embargo, esta visión

agregada puede ocultar diferencias significativas en el comportamiento predictivo entre trenes individuales.

	Modelo	Conjunto	RMSE	MAE	MAPE	R2
1	RLM	Entrenamiento	30061.167	20370.284	12.426413	0.9728572
2	RLM	Prueba	28436.356	22614.309	15.498948	0.9775142
3	XGBoost	Entrenamiento	11986.563	6961.007	3.480087	0.9956845
4	XGBoost	Prueba	46645.008	25337.367	15.681873	0.9394978
5	RF	Entrenamiento	19493.769	12787.319	8.255314	0.9885861
6	RF	Prueba	31450.385	19395.954	14.159596	0.9724949
7	Prophet	Entrenamiento	8214.831	5292.877	3.459238	0.9979731
8	Prophet	Prueba	4544.639	3357.143	2.690524	0.9994257

Tabla 3 Comparativa de resultados por modelo global (RMSE, MAE, MAPE, R^2) (Autoría propia)

Por esta razón, se priorizó la evaluación desagregada por tren (Figura 20), la cual permite identificar con mayor precisión los niveles de ajuste de cada modelo en contextos operativos específicos. Esta comparación individualizada evidenció que, si bien modelos como Prophet alcanzaron altos niveles de R^2 en la mayoría de los trenes, otros algoritmos como Random Forest y XGBoost mostraron desempeño heterogéneo, particularmente en trenes con alta variabilidad o bajo volumen de datos históricos. La evaluación por tren constituye, por tanto, el criterio principal para la selección y análisis de los modelos predictivos en este proyecto.

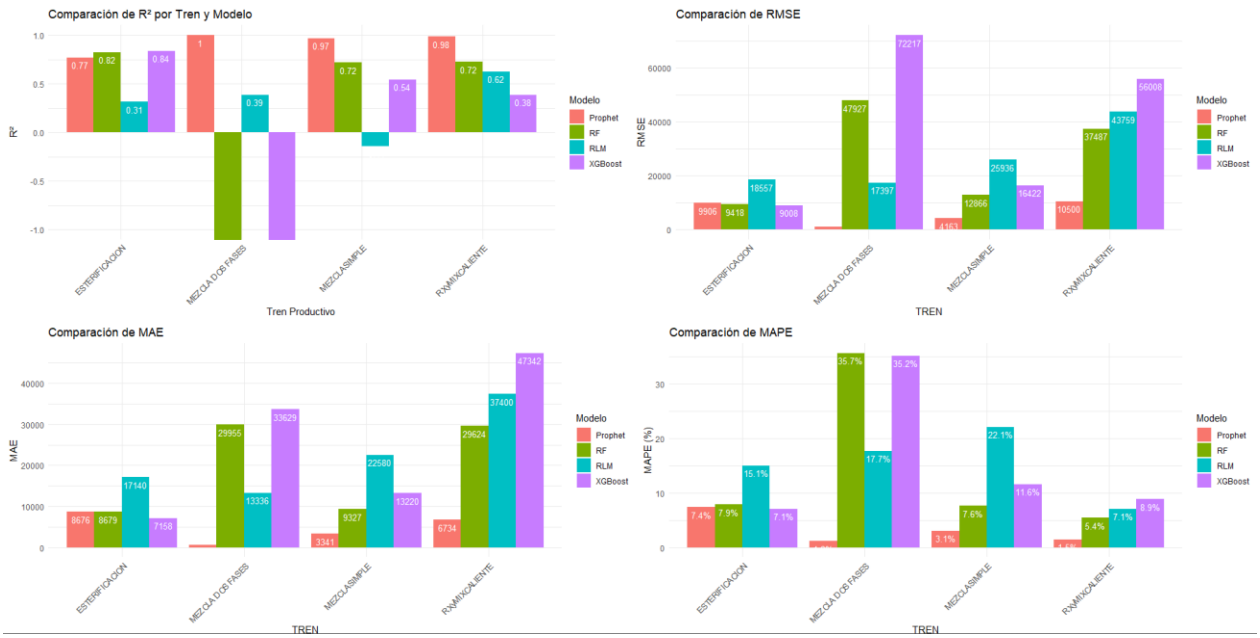

Figura 20 Comparativa de resultados por tren y modelo (RMSE, MAE, MAPE, R^2) (Autoría propia)

La figura 20 presenta una comparación detallada del desempeño de los modelos predictivos desarrollados (Prophet, Random Forest, RLM y XGBoost), evaluados individualmente por tren productivo utilizando las métricas RMSE, MAE, MAPE y R^2.

En términos del coeficiente de determinación (R^2), Prophet se destaca como el modelo más consistente, con valores superiores a 0.77 en todos los trenes y superando 0.97 en mezcla dos fases y Reacciones y mezclas en caliente. XGBoost, por su parte, muestra resultados heterogéneos: si bien presenta un R^2 alto en Esterificación (0.84), su desempeño en mezcla dos fases Fases es negativo, lo cual indica un ajuste deficiente. Random Forest y RLM ofrecen resultados intermedios, con un comportamiento estable en trenes como Mezcla simple.

En cuanto a los errores absolutos (RMSE y MAE), Prophet nuevamente lidera en la mayoría de los trenes, con valores notoriamente más bajos, lo que refleja su capacidad para capturar los patrones de comportamiento mensual. XGBoost y Random Forest, aunque competitivos en algunos trenes, evidencian errores significativamente mayores en contextos con alta variabilidad, como mezcla dos fases.

En la métrica de error porcentual (MAPE), Prophet logra los mejores resultados, manteniéndose por debajo del 5% en tres de los cuatro trenes, lo que evidencia una predicción proporcionalmente más precisa. En contraste, los modelos de árboles presentan valores de MAPE superiores al 15% en trenes con menor estabilidad operativa.

Estos hallazgos sugieren que no existe un único modelo óptimo para todos los trenes, pero Prophet se posiciona como el algoritmo más robusto para contextos con patrones temporales definidos. En cambio, los modelos basados en árboles como RF y XGBoost pueden verse afectados por la variabilidad del entorno batch o por conjuntos de datos con baja representatividad histórica. Esta situación refuerza la necesidad de aplicar un enfoque multialgoritmo por tren, evaluando de forma desagregada para lograr una estimación más fiable de la capacidad alcanzable.

No obstante, es importante aclarar que, si bien Prophet demostró un excelente ajuste en términos de precisión y error, su estructura de modelado basada en series temporales genera una proyección oscilante mes a mes. Por esta razón, no resulta el modelo más apropiado cuando el objetivo es estimar un valor único de capacidad máxima alcanzable por tren, ya que no entrega una línea de referencia fija, sino una serie de valores variables en el tiempo. Para este propósito, modelos como RLM o XGBoost, que permiten derivar un valor proyectado máximo único por tren, resultan más funcionales en la práctica operativa de planificación de capacidad.

5.6 COMPARACIÓN DEL DESEMPEÑO CON Y SIN AJUSTE DE HIPERPARÁMETROS

La Tabla 4 presenta una comparación detallada del desempeño de los modelos XGBoost y Random Forest por tren productivo, tanto bajo configuraciones por defecto como con hiperparámetros ajustados mediante validación cruzada. Si bien en algunos trenes el ajuste produjo mejoras modestas (por ejemplo, en Mezcla simple y RX y mezcla caliente se observó un aumento en R^2 y una reducción de RMSE), en otros casos el ajuste resultó contraproducente. Particularmente, en el tren Esterificación, el modelo ajustado de XGBoost mostró un deterioro significativo en el error (RMSE aumentó de 8.539 a 14.078) y en la capacidad explicativa (R^2 pasó de 0.85 a 0.61).

Esta variabilidad en el impacto del ajuste entre trenes motivó la decisión de mantener los modelos con parámetros por defecto como baseline final. Esta elección permite una configuración homogénea, robusta y replicable en contextos industriales reales, donde las diferencias de volumen de datos entre trenes y la necesidad de mantenimiento del sistema predicen favorecen la simplicidad y estabilidad frente a ganancias marginales en precisión.

Comparación de desempeño por tren y configuración de modelo (XGBoost y Random Forest)

Tren	RMSE XGB (Def)	R^2 XGB (Def)	RMSE XGB (Tuned)	R^2 XGB (Tuned)	RMSE RF (Def)	R^2 RF (Def)	RMSE RF (Tuned)	R^2 RF (Tuned)
ESTERIFICACION	8539.47	0.85	14078.39	0.61	11801.07	0.72	11095.18	0.75
MEZCLA DOS FASES	9003.56	0.84	8522.82	0.85	22741.99	-0.05	19224.81	0.25
MEZCLASIMPLE	17118.23	0.50	14479.83	0.64	10796.55	0.80	17045.86	0.51
RXyMIXCALIENTE	54025.25	0.43	40710.32	0.68	40061.67	0.69	38087.16	0.72

Tabla 4 Comparación de desempeño por tren y configuración de modelo (XGBoost y Random Forest) (Autoría propia)

5.7 ANALISIS POR TREN PRODUCTIVO

A partir de la evaluación desagregada por tren, se identificaron patrones diferenciados en el ajuste y desempeño de los modelos implementados. Si bien Prophet ofreció altos niveles de precisión en la mayoría de los casos, se tuvieron en cuenta tanto la precisión estadística como la aplicabilidad práctica al objetivo del estudio: estimar la capacidad máxima alcanzable por tren.

Esterificación: Prophet y XGBoost presentaron los mejores desempeños, alcanzando valores de R^2 de 0.77 y 0.84 respectivamente. La capacidad proyectada en este tren estuvo altamente explicada por la disponibilidad operativa mensual y las horas efectivas de proceso, lo que favoreció a modelos sensibles a patrones históricos y no lineales.

Mezcla dos fases: Prophet logró un ajuste sobresaliente con un R^2 de 0.997, capturando con precisión la estacionalidad y los picos de producción. No obstante, los modelos basados en árboles (XGBoost y Random Forest) mostraron un desempeño negativo en R^2, atribuible a la alta

variabilidad operativa y dispersión de registros en este tren. Esto sugiere que modelos lineales o de tendencia suavizada tienen mayor estabilidad aquí.

Mezcla simple: Prophet ($R^2 = 0.969$) y Random Forest ($R^2 = 0.719$) fueron los algoritmos con mejor comportamiento. Este tren presenta procesos menos complejos y más homogéneos, lo que permitió un ajuste adecuado tanto para modelos secuenciales como basados en reglas.

Reacciones y mezclas en caliente: Prophet nuevamente obtuvo el mejor desempeño ($R^2 = 0.985$), mostrando gran capacidad para modelar la tendencia general. Sin embargo, los modelos tradicionales (RLM, XGBoost, RF) enfrentaron mayores dificultades para capturar la variabilidad en tiempos de reacción y la dispersión de los lotes. A pesar de su buen ajuste, el carácter oscilante de las proyecciones de Prophet limita su utilidad como estimador único de capacidad máxima, lo cual debe considerarse en la decisión final del modelo aplicable.

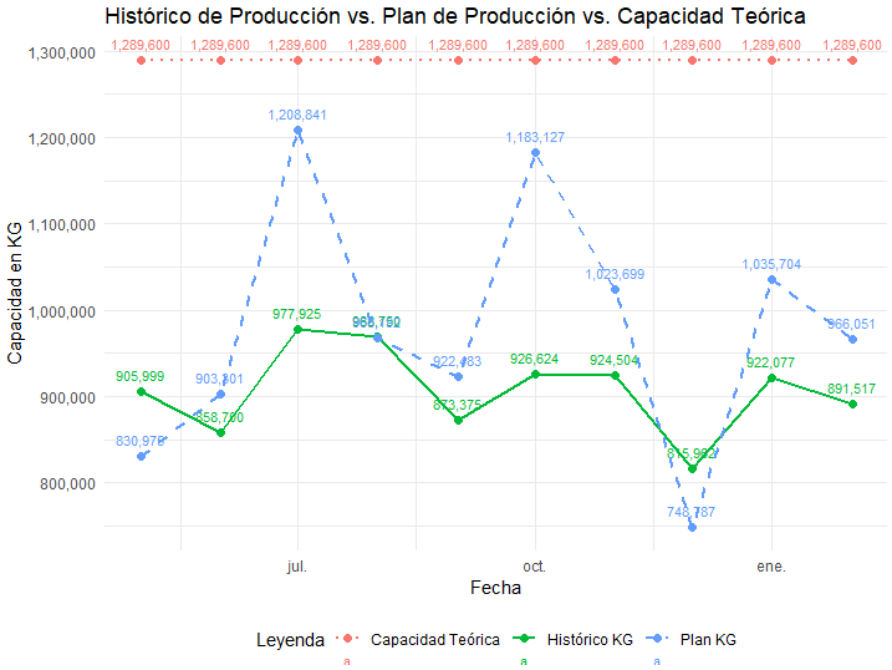

Figura 21 Histórico de producción vs plan vs capacidad teorica por tren (Autoría propia)

Esta figura permite visualizar la brecha entre tres elementos clave de la planificación operativa: la producción histórica efectiva (línea verde), el plan mensual de producción (línea azul) y la capacidad teórica máxima definida por la organización (línea roja punteada). Se observa que el plan supera regularmente la producción alcanzada. Esta comparación refuerza la necesidad de contar con estimaciones realistas de capacidad alcanzable para mejorar la alineación entre plan y realidad operativa.

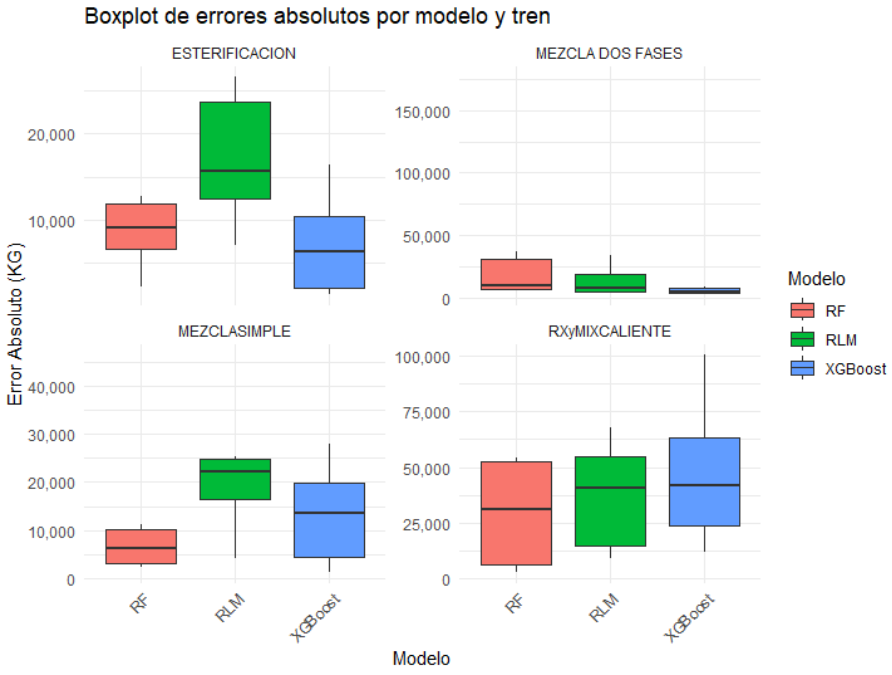

Figura 22 Boxplot de errores absolutos por modelo y tren (Autoría propia)

La figura muestra la distribución del error absoluto en kilogramos por modelo (RF, RLM, XGBoost) y por tren. Random Forest presenta una dispersión más controlada en la mayoría de los casos, mientras que RLM muestra mayor variabilidad, especialmente en esterificación. en mezcla dos fases, XGBoost logra un error muy bajo y consistente, mientras que en reacciones y mezclas en caliente todos los modelos presentan mayor dispersión, evidenciando la complejidad de ese tren. Este análisis complementa las métricas agregadas y permite observar no solo el promedio de error, sino también su estabilidad y distribución.

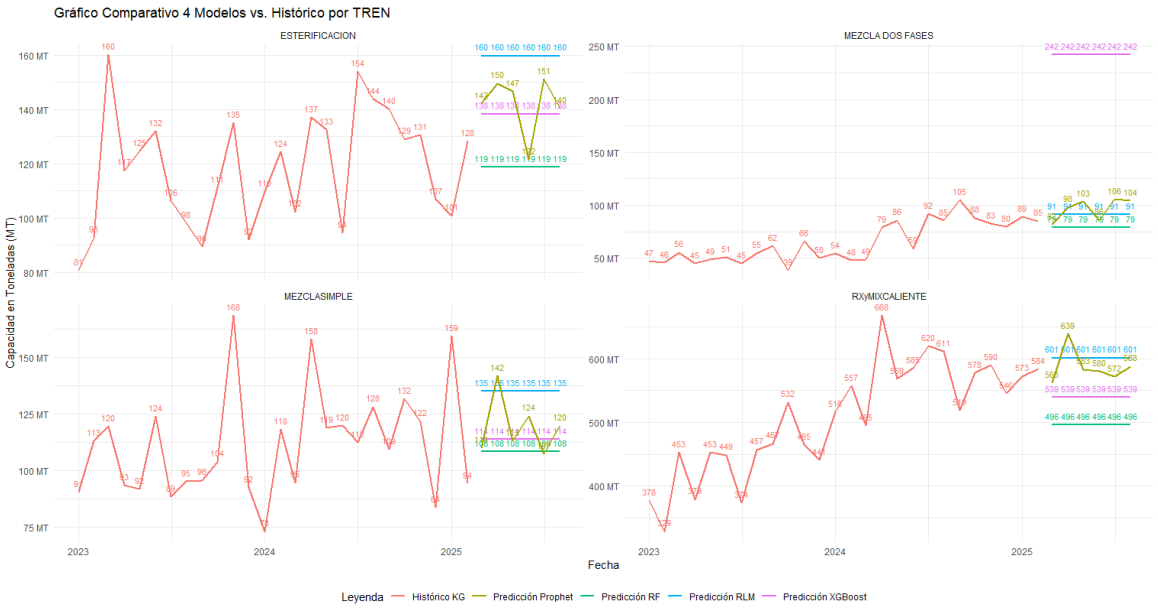

Figura 23 Comparación de modelos Prophet, RF, XGB, RLM vs histórico (Autoría propia)

Aquí se comparan directamente las predicciones finales de los cuatro modelos frente al histórico real por tren. Se puede observar cómo RLM y XGBoost tienden a generar una proyección estable de capacidad máxima, alineada con el objetivo del estudio. En contraste, Prophet presenta proyecciones fluctuantes mes a mes, lo cual, aunque útil para seguimiento temporal, no es ideal para fijar un límite máximo alcanzable. Esta visualización consolida la decisión metodológica de priorizar modelos que entregan un valor único y proyectable de capacidad máxima mensual por tren, facilitando su uso en planeación estratégica.

5.8 LIMITACIONES Y OPORTUNIDADES FUTURAS

Aunque el enfoque propuesto logró resultados coherentes con el comportamiento real de los trenes productivos, es importante reconocer ciertas limitaciones que pueden afectar la precisión y generalización de los modelos desarrollados.

En primer lugar, algunos trenes, especialmente Reacciones y mezclas en caliente, presentaron un volumen limitado de datos históricos, lo que afecta la robustez estadística y amplifica la sensibilidad ante valores atípicos. Además, los primeros meses del registro carecían de información completa sobre OEE, lo que obligó a imputar la disponibilidad con promedios históricos. Aunque esta estrategia permitió continuar el análisis, también introduce un sesgo al homogeneizar periodos que podrían haber presentado comportamientos anómalos.

Por otro lado, aunque Prophet mostró un excelente desempeño estadístico en la mayoría de los trenes, modelos como XGBoost y Random Forest evidenciaron dificultades en líneas con alta variabilidad, como mezcla dos fases, resaltando la necesidad de estrategias diferenciadas por contexto operativo.

Otro aspecto relevante es que el modelo se construyó exclusivamente sobre variables internas del proceso, sin incluir factores externos que también condicionan la capacidad productiva, como la demanda programada, los turnos laborales, la disponibilidad de materias primas o restricciones logísticas. Incorporar estos elementos en el futuro permitiría enriquecer los modelos y acercarlos a un enfoque más holístico para la toma de decisiones.

Finalmente, el periodo de análisis (2023–2025) fue adecuado para el desarrollo inicial, pero ampliar la ventana temporal permitiría aplicar técnicas más sólidas de análisis estacional o validación a largo plazo.

5.9 CONCLUSIONES DEL MODELO PREDICTIVO

El desarrollo del modelo predictivo permitió estimar de forma precisa y contextualizada la capacidad mensual máxima alcanzable por tren productivo, integrando datos operativos históricos con técnicas de aprendizaje supervisado. Se validó que variables como disponibilidad operativa, horas efectivas y ocupación relativa a la capacidad teórica son los principales predictores de desempeño, justificando su inclusión en el modelo.

Aunque Prophet obtuvo altos valores de R^2 en todos los trenes, su carácter oscilante y su diseño enfocado en la detección de tendencias estacionales lo hacen menos adecuado para el objetivo específico de este estudio, que es definir un valor único de capacidad técnica máxima bajo condiciones actuales. Por ello, se priorizaron modelos como XGBoost, Random Forest y RLM, que permiten estimaciones planas, concretas y explicables en función de variables internas.

Entre ellos, XGBoost se destacó en trenes con alta estabilidad operativa como Esterificación, mientras que Random Forest mostró mayor robustez ante variabilidad en trenes como Mezcla simple y Reacciones y mezclas en caliente. RLM, aunque con menor precisión global, fue útil por su transparencia y por permitir una lectura clara del peso estadístico de cada predictor.

Desde una perspectiva aplicada, los modelos desarrollados ofrecen un insumo valioso para la planificación operativa, ya que permiten anticipar la capacidad técnica disponible de forma más realista, complementar la planificación basada en demanda y detectar posibles cuellos de botella. Además, proporcionan una herramienta explicativa para comprender por qué ciertos trenes logran un mejor desempeño que otros, sentando las bases para futuros modelos prescriptivos que orienten decisiones sobre asignación de recursos y cierre de brechas entre lo teórico y lo real.

El código completo empleado para la construcción, ajuste y validación de los modelos se encuentra disponible en el Anexo 1.

Una vez seleccionados los modelos con mejor desempeño, se procedió a contrastar sus predicciones con la capacidad teórica definida por la organización para cada tren productivo. Esta comparación constituye un insumo clave para identificar brechas y orientar decisiones estratégicas. En el siguiente capítulo se presenta este análisis comparativo.

6 ANALISIS DE CAPACIDADES PROYECTADAS Y CAPACIDADES TEORICAS

Una vez desarrollados, ajustados y validados los modelos de predicción por tren, el siguiente paso en el proyecto consistió en comparar los resultados proyectados por los modelos frente a la capacidad nominal o teórica definida para cada tren productivo. Esta comparación permite identificar brechas entre lo que el modelo estima como alcanzable (en función de la disponibilidad, ocupación y otras condiciones históricas) y lo que la organización espera lograr bajo condiciones ideales.

6.1 CAPACIDAD TEÓRICA VS. PROYECTADA POR TREN

La capacidad teórica mensual fue extraída del estándar técnico de producción de cada tren, que define cuántos kilogramos podrían fabricarse mensualmente asumiendo un 100% de disponibilidad y eficiencia. Esta cifra fue comparada con la capacidad proyectada por los modelos de mayor desempeño (en la mayoría de los casos, XGBoost), a partir de datos históricos y con escenarios de entrada promedio.

Para estandarizar la comparación, se tomó como referencia el valor medio mensual proyectado por los modelos para el periodo entre agosto de 2024 y febrero de 2025. Estos valores fueron contrastados contra la capacidad teórica oficial de cada tren, y se calculó el porcentaje de cumplimiento o desviación. Esta métrica permite visualizar en qué trenes se está aprovechando mejor la capacidad instalada y en cuáles hay margen de mejora.

Tren Productivo	Capacidad Teórica (MT)	Capacidad Proyectada (MT)	Modelo Escogido	Gap (%)
ESTERIFICACION	176.0	138.34256	XGBoost	21.4%
MEZCLA DOS FASES	166.4	91.45615	RLM	45%
MEZCLASIMPLE	124.8	108.25395	RF	13.3%
RXyMIXCALIENTE	822.4	496.27625	RF	39.7%

Tabla 5 Resumen de capacidad teórica vs. proyectada por tren (Autoría propia)

Se resume los resultados por tren, mostrando la capacidad teórica, la capacidad proyectada por el modelo, el algoritmo seleccionado y la brecha porcentual entre ambas.

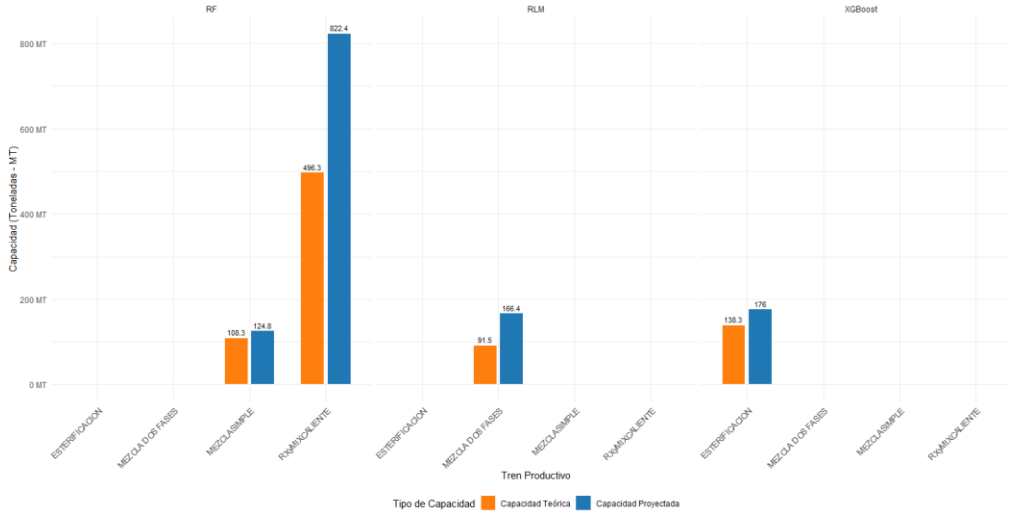

Figura 24 Gráfico de barras por tren capacidad teórica y capacidad proyectada (Autoría propia)

6.2 RESULTADOS POR TREN

Esterificación: El modelo XGBoost proyectó una capacidad mensual de 138.3 toneladas, equivalente al 78.6% de la capacidad teórica. La brecha del 21.4% puede atribuirse a pérdidas por cambios de lote, paradas programadas y tiempos muertos.

Mezcla dos fases: La RLM fue el modelo más estable para este tren. Proyectó 91.5 toneladas frente a un estándar de 166.4 toneladas, con una brecha de 45%, la más alta entre los trenes. Esto indica una subutilización estructural, probablemente ligada a disponibilidad restringida o baja asignación de lotes.

Mezcla simple: Con un rendimiento del 86.7% frente a su capacidad nominal, este tren mostró el menor gap (13.3%). El modelo Random Forest estimó una producción mensual promedio de 108.2 toneladas. Este resultado sugiere un aprovechamiento eficiente, aunque con oportunidad de mejorar tiempos de preparación y secuencia.

Reacciones y mezclas en caliente: Este tren presentó una brecha del 39.7%, con una capacidad proyectada de 496.3 toneladas frente a un estándar de 822.4 toneladas. Aunque Random Forest fue el modelo más adecuado, su desempeño refleja la alta variabilidad operativa y utilización intermitente de esta línea.

6.3 INTERPRETACIÓN DE RESULTADOS Y RECOMENDACIONES

Este análisis de brechas permite identificar prioridades de mejora en términos de eficiencia operativa. En trenes con alta desviación negativa respecto a la capacidad teórica, se recomienda profundizar en estudios de tiempos no productivos, paradas no programadas, y revisión del plan maestro de producción (MPS) para verificar si las metas asignadas están en línea con las restricciones reales. También es recomendable revisar la política de asignación de lotes a trenes, priorizando aquellos con menor desviación cuando sea técnicamente posible.

Por otro lado, los trenes con bajo margen de brecha validan que las condiciones actuales permiten un uso casi óptimo de la infraestructura disponible. En estos casos, se sugiere sostener estrategias actuales de mantenimiento, programación y asignación, e incluso considerar estos trenes como referencia para planes de estandarización en los otros.

6.4 ANÁLISIS DE BRECHAS

La comparación entre la capacidad proyectada por modelos y la capacidad teórica establecida por la organización es un paso esencial para traducir el análisis predictivo en acciones concretas de mejora. Este ejercicio permite contextualizar las predicciones dentro de la realidad operativa esperada y ofrece una herramienta cuantitativa para definir prioridades y justificar decisiones de planificación, inversión o reconfiguración operativa. El análisis también abre la puerta a simulaciones futuras, donde se puedan modelar escenarios hipotéticos de mejora (por ejemplo, un aumento en disponibilidad o eficiencia), y medir su impacto sobre la brecha observada.

La identificación de brechas y la proyección de capacidades alcanzables genera valor agregado cuando se convierte en herramienta fundamental para la toma de decisiones. Por ello, en el capítulo siguiente se describe el diseño de dashboards operativos y estratégicos que permiten integrar estos hallazgos en la toma de decisiones diaria por parte de los equipos de manufactura y planeación.

Si bien los modelos predictivos permiten estimar la capacidad alcanzable, su verdadero valor se concreta al integrarlos en el sistema de gestión de la planta. En el siguiente capítulo se describe cómo se diseñaron tableros de control operativos y estratégicos que permiten convertir estas predicciones en herramientas para la toma de decisiones en tiempo real.

7 DISEÑO E IMPLEMENTACIÓN DE TABLEROS DE CONTROL

Uno de los objetivos clave de este proyecto, además del desarrollo del modelo predictivo, fue asegurar que los resultados pudieran ser utilizados de forma práctica y continua por los equipos de planeación y manufactura. Para ello, se diseñaron dashboards interactivos que permiten monitorear la capacidad proyectada por tren, contrastarla con la capacidad teórica y dar seguimiento a las brechas operativas en tiempo real.

Estos dashboards fueron desarrollados en Power BI, integrando directamente los resultados de los modelos entrenados en R y los datos operativos históricos almacenados en Google Sheets y bases internas. Su diseño se orientó a facilitar la toma de decisiones tácticas y estratégicas mediante visualizaciones intuitivas, indicadores clave (KPIs) y filtros interactivos por periodo, tren productivo o tipo de capacidad.

A lo largo de este capítulo, se presentan los principales componentes del dashboard desarrollado, destacando cómo estos permiten transformar el análisis predictivo en una herramienta de apoyo directo para la operación diaria y la planeación mensual de la planta.

7.1 FLUJO DE DATOS Y ACTUALIZACIÓN DE TABLEROS

Aunque ambos tableros tienen como fuente de origen los registros capturados en Google Forms y almacenados en Google Sheets, su flujo de procesamiento y nivel de actualización presentan diferencias significativas:

Tablero Operativo en Looker Studio:

- Flujo de datos: Google Forms → Google Sheets → Looker Studio.
- Actualización: Actualización automática en tiempo real, sin intervención manual.
- Propósito: Supervisión operativa diaria, control en piso de los estados de los reactores, y validación inmediata de la captura de datos.

Tablero Estratégico en Power BI:

- Flujo de datos: Google Forms → Google Sheets → R (API) → Modelado y cálculos → Exportación manual → Power BI.
- Actualización: Manual, a partir de la ejecución de scripts de tratamiento y modelado en R.
- Propósito: Análisis predictivo estratégico de la capacidad instalada, identificación de brechas productivas, soporte a la toma de decisiones de planeación.

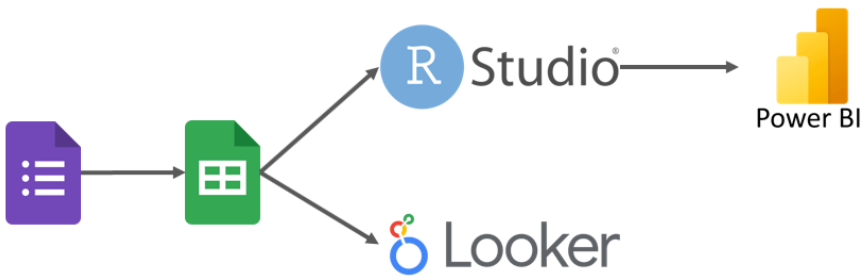

Figura 25 Flujo de datos para los tableros de control operativo y estratégico (Autoría propia)

La figura muestra el flujo de integración de datos desde su captura en planta hasta su visualización para la toma de decisiones. Los registros operativos se recolectan en tiempo real a través de formularios digitales (Google Forms), los cuales alimentan automáticamente una base central en Google Sheets. Desde allí, los datos siguen dos rutas complementarias:

Ruta analítica: los datos son procesados y modelados en RStudio, donde se calculan predicciones de capacidad y se transforman las variables. Los resultados se exportan a Power BI para construir dashboards estratégicos.

Ruta operativa: los datos se visualizan directamente en Looker Studio (anteriormente Data Studio), permitiendo un monitoreo ágil y cotidiano de KPIs como OEE, ocupación o volumen producido.

Este esquema garantiza una solución flexible, integrada y accesible, donde las áreas operativas pueden consultar datos actualizados sin depender de procesos manuales.

7.2 TABLERO OPERATIVO EN LOOKER STUDIO

El tablero operativo fue diseñado para ofrecer una visualización en tiempo real del estado de los procesos productivos en las distintas plantas de la organización. Gracias a la conexión directa entre los formularios digitales de captura, las hojas de registro (Google Sheets) y el dashboard en Looker Studio, se logra una visualización dinámica, actualizada y accesible de:

- El estado actual de cada reactor o equipo.
- El producto y lote en proceso.
- La etapa específica del proceso (inicio, ajuste, reacción, descarga, parada).
- La hora y fecha de registro de cada etapa.

Este tablero permite que operarios, supervisores, líderes de turno, gerentes de planta y directores

accedan a la información desde diferentes dispositivos (pantallas de planta, computadores, tablets o celulares), facilitando la coordinación en tiempo real y el monitoreo continuo de la operación.

Además, cumple un rol crítico en la calidad de los datos empleados en los modelos predictivos, ya que la visibilidad inmediata promueve la correcta ejecución de los registros de cada etapa. Esto mejora la trazabilidad de los datos, incrementa su confiabilidad y asegura que las proyecciones de capacidad futuras se basen en información precisa y oportuna.

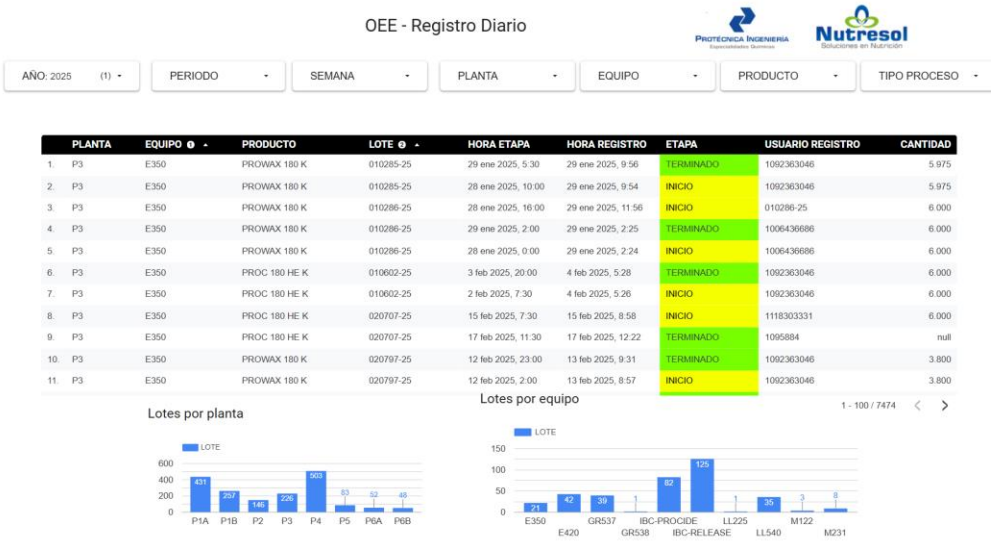

Figura 26 Tablero operativo en Looker Studio – Registro diario detallado de lotes, etapas y reactores (Autoría propia)

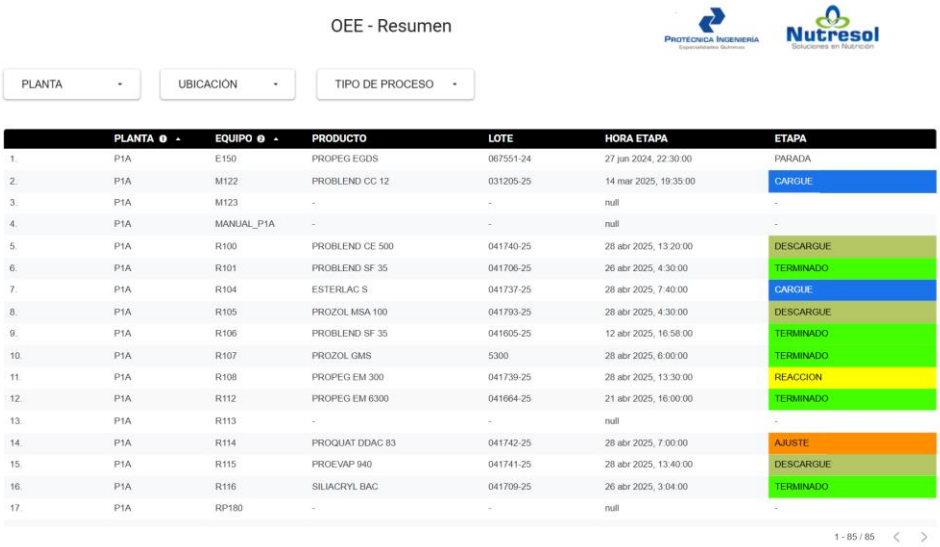

Figura 27 Tablero operativo en Looker Studio – Estado actual de reactores por planta (Autoría propia)

7.3 INTEGRACIÓN DE TABLEROS PARA LA TOMA DE DECISIONES

La construcción de tableros de control diferenciados según el nivel de uso operativo o estratégico garantiza que la información capturada en planta no solo sea utilizada para supervisar el día a día de los procesos, sino también para alimentar sistemas de análisis predictivo que respalden la planificación y la optimización de recursos a nivel organizacional.

Mientras el tablero operativo en Looker Studio asegura la captura oportuna y precisa de los eventos de fabricación en tiempo real, el tablero estratégico en Power BI consolida la información procesada y modelada, proporcionando una herramienta analítica robusta para evaluar tendencias, brechas de capacidad y oportunidades de mejora.

De esta forma, ambos tableros cumplen funciones complementarias en el ecosistema de gestión de datos de la organización:

- Looker Studio como soporte a la operación diaria en planta.
- Power BI como soporte a la planeación táctica y estratégica.

La siguiente sección describe en detalle el diseño y funcionalidad del tablero estratégico en Power BI.

7.4 TABLERO ESTRATÉGICO EN POWER BI

Paralelamente, se diseñó la estructura conceptual de un tablero estratégico en Power BI orientado a consolidar la información procesada y modelada en R. Este tablero incluirá visualizaciones como:

- Predicción de capacidad mensual alcanzable por tren.
- Comparaciones entre la capacidad real proyectada y la capacidad nominal definida.
- Análisis de brechas operativas por tren y planta.
- Tendencias históricas de producción y ocupación.

Aunque al momento de esta entrega el tablero se encuentra en fase de implementación, su diseño ya permite consultar predicciones desagregadas por planta, tren y periodo, con filtros interactivos y visualización clara de los resultados.

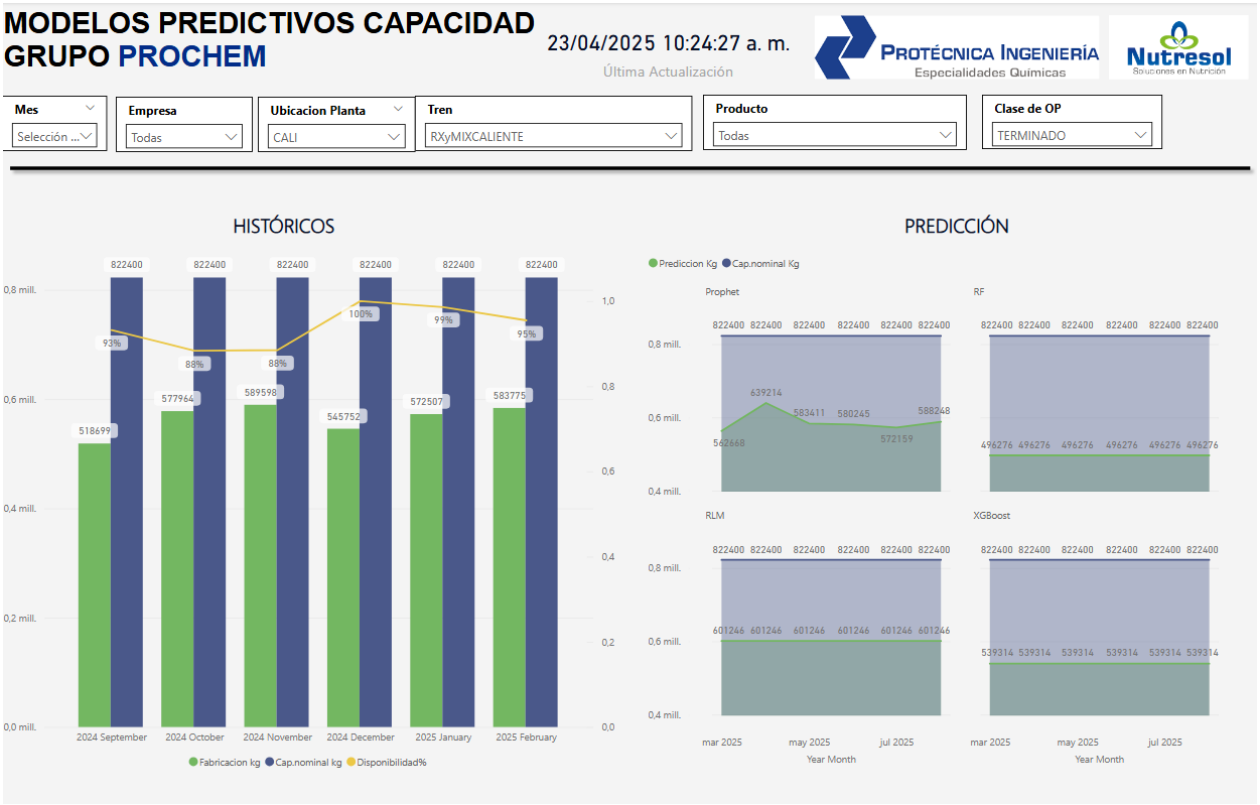

Figura 28 Tablero estratégico en Power BI – Modelo predictivo vs capacidades teóricas vs ejecución (Autoría propia)

7.5 PROPUESTA DE MEJORA FUTURA

Actualmente, el tablero operativo en Looker Studio opera en tiempo real gracias a la conexión directa con los registros en Google Sheets. Sin embargo, el tablero estratégico en Power BI depende de un proceso manual de ejecución de scripts y exportación de resultados.

Una mejora futura relevante sería la automatización del flujo de actualización del tablero estratégico, mediante:

- La programación de tareas automáticas de ejecución de scripts en R (cron jobs o Programador de tareas).
- La implementación de flujos de actualización en Power BI Service (mediante Power Automate o programación de data sets).
- La migración del tratamiento de datos a plataformas como Microsoft Fabric.

Esta automatización permitiría consolidar un ecosistema de toma de decisiones basado en datos confiables, actualizados periódicamente, y con mínima intervención humana, fortaleciendo la agilidad operativa y estratégica de la organización.

8 CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

El presente proyecto tuvo como finalidad desarrollar un sistema integral de captura, análisis predictivo y visualización de la capacidad productiva de trenes de especialidades químicas. A partir de una situación inicial caracterizada por una limitada disponibilidad de registros históricos estructurados, se diseñaron mecanismos de captura de datos en tiempo real, combinados con procesos de limpieza, normalización y tratamiento avanzado de datos.

La etapa de modelado predictivo permitió construir algoritmos adaptados a las características operativas de cada tren productivo, utilizando técnicas de Regresión Lineal Múltiple, Random Forest, XGBoost y Prophet. Los modelos lograron explicar una proporción significativa de la variabilidad observada en la capacidad mensual alcanzada, con desempeños diferenciados entre trenes según su nivel de estabilidad operativa y volumen histórico de datos disponible.

El análisis de brechas entre la capacidad proyectada por los modelos y la capacidad nominal definida permitió identificar trenes con niveles adecuados de aprovechamiento, así como otros con oportunidades claras de mejora. Estos resultados aportan una primera aproximación cuantitativa al entendimiento de los límites operativos reales de la planta, aunque deben ser interpretados considerando las limitaciones de la calidad y volumen de los datos disponibles.

La implementación de tableros de control diferenciados, uno operativo en Looker Studio y otro estratégico en Power BI, representa un avance importante en la democratización del acceso a la información, permitiendo a distintos niveles de la organización supervisar, analizar y actuar sobre los datos de manera oportuna.

Si bien los resultados obtenidos son alentadores y demuestran la aplicabilidad práctica del enfoque desarrollado, es importante reconocer que su impacto efectivo dependerá de la evolución de la cultura de captura de datos en planta, de la sostenibilidad de los procesos de actualización de los modelos, y de la incorporación progresiva de variables externas que puedan enriquecer las predicciones.

Por lo tanto, los hallazgos y herramientas aquí presentados deben entenderse como un punto de partida para un proceso continuo de mejora en la gestión de la capacidad productiva basada en datos, y no como una solución cerrada o definitiva.

Es importante considerar que los modelos predictivos desarrollados en este proyecto reflejan la capacidad alcanzable bajo las condiciones históricas observadas de infraestructura, equipos y recursos disponibles. En caso de introducir cambios significativos en el diseño e infraestructura de planta, tales como la adquisición de nuevos reactores, ampliaciones de planta o modificaciones sustanciales en la disponibilidad de personal operativo, los modelos actuales podrían perder validez, ya que estarían basados en patrones de comportamiento que no incluirían dichas variaciones. Por tanto, se recomienda que cualquier modificación relevante en la capacidad

instalada sea acompañada de un proceso de actualización de los modelos, incorporando los nuevos registros operativos para garantizar la fiabilidad de las proyecciones.

8.2 TRABAJOS FUTUROS

A partir del desarrollo y resultados obtenidos en este proyecto, se identifican varias líneas de trabajo que podrían fortalecer, expandir o actualizar los modelos y herramientas implementadas:

Automatización de la actualización de tableros estratégicos: Actualmente, el flujo de actualización del tablero en Power BI depende de la ejecución manual de scripts de procesamiento en R. Una mejora prioritaria sería la automatización de este proceso, mediante la programación de tareas periódicas de ejecución de scripts (cron jobs o programador de tareas), o la implementación de flujos de integración automática hacia Power BI Service o plataformas de orquestación en la nube.

Incorporación de nuevas variables explicativas: Aunque los modelos actuales se basan en variables operativas como disponibilidad, horas efectivas y eficiencia, la inclusión de nuevas variables como calidad de materias primas, número de cambios de producto, tiempos de limpieza, o información de mantenimiento preventivo podría enriquecer los modelos predictivos y mejorar su capacidad explicativa.

Actualización de modelos ante cambios de infraestructura: Dado que los modelos predictivos fueron entrenados bajo las condiciones actuales de equipos e infraestructura, cualquier incorporación de nuevos reactores, ampliaciones de planta, o cambios significativos en la capacidad instalada requerirá un proceso formal de actualización o reentrenamiento de los modelos, incorporando nuevos registros que reflejen las nuevas condiciones operativas.

Desarrollo de modelos adaptativos: Como extensión futura, podría explorarse el desarrollo de modelos de Machine Learning adaptativos o en línea, que permitan ajustar los parámetros de predicción a medida que se reciben nuevos datos, reduciendo así la necesidad de reentrenamientos manuales periódicos.

Simulación de escenarios de mejora de eficiencia: Aunque no fue abordado en este proyecto, la simulación de escenarios hipotéticos de mejora (por ejemplo, reducción de tiempos muertos, aumento de disponibilidad) podría ser una herramienta valiosa para evaluar el impacto potencial de iniciativas de optimización operativa sobre la capacidad proyectada.

Integración de indicadores de OEE: Adicionalmente, se sugiere explorar la integración de indicadores de eficiencia global de los equipos (OEE) como variable de análisis complementaria, para fortalecer la conexión entre desempeño de planta y capacidad proyectada.

9 REFERENCIAS BIBLIOGRÁFICAS

[1] J. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, O'Reilly Media, Sebastopol, CA, USA, 2013.

[2] C. W. Dawson, *Projects in Data Science and Machine Learning*, Springer, Cham, Switzerland, 2021.

[3] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, San Francisco, CA, USA, 1999.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[5] W. J. Stevenson, *Operations Management*, 14th ed., McGraw-Hill Education, New York, NY, USA, 2020.

[6] T. E. Vollmann, W. L. Berry, D. C. Whybark, and F. R. Jacobs, *Manufacturing Planning and Control for Supply Chain Management*, McGraw-Hill, New York, NY, USA, 2005.

[7] M. Pinedo, *Scheduling: Theory, Algorithms, and Systems*, Springer, New York, NY, USA, 2016.

[8] R. Carbonneau, K. Laframboise, and R. Vahidov, "Application of machine learning techniques for supply chain demand forecasting," *European Journal of Operational Research*, vol. 184, no. 3, pp. 1140–1154, 2008.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, San Francisco, CA, USA, 2011.

[10] A. Montgomery and C. Jennings, *Introduction to Linear Regression Analysis*, 5th ed., Wiley, Hoboken, NJ, USA, 2011.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, 2016, pp. 785–794.

[13] S. J. Taylor and B. Letham, "Forecasting at Scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed., Springer, New York, NY, USA, 2021.

[15] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, New York, NY, USA, 2013.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[17] P. Rouse, *Business Intelligence Techniques: A Perspective from Accounting and Finance*, Springer, London, UK, 2014.

[18] M. Sherman, *Business Intelligence Guidebook: From Data Integration to Analytics*, Morgan Kaufmann, Burlington, MA, USA, 2014.

[19] Z. Mouhib, M. Gallab, S. Merzouk, A. Soulhi, and B. Elbhiri, "Towards a generic framework of OEE monitoring for driving effectiveness in digitalization era," *Procedia Computer Science*, vol. 232, pp. 2508–2520, 2024.

[20] Y. Ramírez Rodríguez, "Diseño e implementación del sistema de productividad y mejoramiento OEE (Overall Effectiveness Equipment) en las líneas de producción de la compañía OLEOFLORES S.A.S.", Tesis de pregrado, Universidad del Magdalena, 2018. [Online]. Available: https://repositorio.unimagdalena.edu.co/items/4700bb27-549a-4cf6-944e-e829a4896293

[21] M. Díaz, "Diseño e implementación de un sistema BI para análisis de indicadores de eficiencia en una empresa de manufactura de equipos de seguridad," Tesis de maestría, Universidad de los Andes, 2021. [Online]. Available: https://hdl.handle.net/1992/50489

10 ANEXOS

Anexo 1. Código fuente en R para desarrollo de modelo predictivo

El código completo desarrollado en R para la captura, procesamiento, modelado y validación de los datos puede consultarse en el siguiente repositorio digital:
https://github.com/danielfduarteq/ModelosPredictivos_ProyectoAplicado_MAESTRIA_DS

Nota: El repositorio incluye los scripts utilizados para la conexión a Google Sheets, tratamiento de datos, entrenamiento de modelos (Random Forest, XGBoost, Prophet y RLM), validación cruzada y exportación de resultados para Power BI.