Escuela de Administración

Maestría en Business Analytics

Transformación del Acopio de Leche en una Compañía Láctea mediante Analítica Descriptiva y Predictiva para la Optimización del Proceso Productivo

Presentado por:

Joan Sebastian Henao Fontecha

Bogotá, D.C. 17 de noviembre de 2025

Escuela de Administración

Maestría en Business Analytics

Transformación del Acopio de Leche en una Compañía Láctea mediante Analítica Descriptiva y Predictiva para la Optimización del Proceso Productivo

Presentado por:

Joan Sebastian Henao Fontecha

Bajo la dirección de:

Yudy Castaño Aristizabal

Bogotá, D.C. 17 de noviembre de 2025

Contenido

Agradecimientos

Agradezco a Dios por acompañarme en cada paso de este camino, por darme el entendimiento y la claridad para seguir adelante.

A mi familia, que ha sido mi mayor soporte desde siempre. A mis amigos, por estar ahí con su apoyo sincero y su buena vibra cuando más lo necesitaba. Y a mi novia, por su paciencia, cariño y ese apoyo moral que tantas veces me sostuvo.

A la Fundación Tomas Rueda Vargas, gracias por creer en mí y por el apoyo económico que hizo mejor transitable este proceso.

A la empresa donde trabajo y, en especial, a mi jefe, gracias por la comprensión, el tiempo brindado y la flexibilidad que me permitieron equilibrar mis responsabilidades laborales con este proyecto académico.

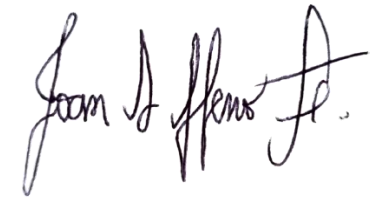
Finalmente, a la Universidad del Rosario, gracias por abrirme sus puertas y ofrecer un espacio donde el conocimiento se convierte en una herramienta de crecimiento personal. Todo lo aprendido aquí me ha enriquecido de formas que llevaré conmigo para siempre.

Joan Sebastian Henao Fontecha

Declaración de originalidad y autonomía

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi propia cuenta y que, por lo tanto, su contenido es original.

Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

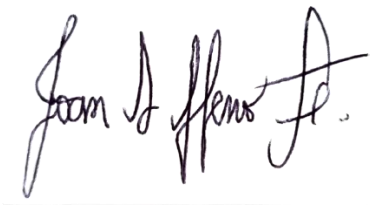


Joan Sebastian Henao Fontecha

Firmado en Bogotá, D.C. el 17 de noviembre de 2025

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de su autor. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Joan Sebastian Henao Fontecha

Firmado en Bogotá, D.C. el 17 de noviembre de 2025

Lista de tablas

Lista de Figuras

Glosario

- **Acidez:** Indicador fisicoquímico que refleja el estado de frescura y la actividad microbiológica de la leche.
- **Acopio:** Proceso de reunir, recibir y almacenar materias primas (en este caso, leche) provenientes de diversos proveedores ganaderos.
- **Cadena de frío:** Sistema de conservación a baja temperatura que evita el deterioro microbiológico y mantiene la calidad del producto.
- **Centros de acopio:** Infraestructura destinada a recibir, almacenar y conservar la leche antes de su procesamiento industrial.
- **EBITDA:** Indicador financiero que mide la rentabilidad operativa antes de intereses, impuestos, depreciaciones y amortizaciones.
- **Estacionalidad:** Variación cíclica en la producción y calidad de la leche asociada a cambios climáticos y disponibilidad de forraje.
- **Grasa láctea: F**racción lipídica de la leche compuesta principalmente por triacilglicéridos; clave para sabor, textura y rendimiento.
- **Incentivos por calidad:** Esquema de pago diferencial basado en atributos fisicoquímicos o microbiológicos para estimular mejores prácticas productivas.
- **Leche cruda:** Materia prima perecedera compuesta por agua y sólidos (grasa, proteína, lactosa, vitaminas y minerales).
- **Perfil composicional:** Conjunto de atributos fisicoquímicos que describen la calidad de la leche, como grasa, proteína, SNG y ST.
- **Proteína láctea:** Componente nitrogenado de la leche (caseínas y proteínas de suero) esencial para la elaboración de productos lácteos.

- **Rendimiento industrial:** Cantidad de producto final obtenido por unidad de leche, directamente influenciada por su composición y calidad.
- **Sólidos no grasos (SNG):** Fracción de la leche conformada por proteínas, lactosa y minerales, excluyendo la grasa.
- **Sólidos totales (ST):** Porción no acuosa de la leche que integra grasa, proteínas, lactosa y minerales.

Resumen Ejecutivo

El presente estudio de caso aborda la necesidad de transformar la toma de decisiones basadas en datos para mejorar la planificación productiva y el rendimiento industrial asociados al acopio de leche cruda en una compañía láctea. Para ello, se aplicó la metodología CRISP-DM mediante un enfoque que integra análisis descriptivo y predictivo. En la fase descriptiva se utilizaron técnicas de PCA y K-means, identificando tres segmentos de leche diferenciados por su composición (Leche Premium, Leche para Leche y Leche Estándar) lo que permite asignar estratégicamente la materia prima de mayor calidad a procesos de alto valor, haciendo más eficiente el proceso. En la fase predictiva se evaluaron modelos ARIMA y SARIMAX, seleccionándose este último por su mayor precisión con parámetros (1,0,2)x(1,1,1,30), gracias a la inclusión de la variable exógena Grasa y su capacidad para capturar la estacionalidad mensual; pese a la volatilidad diaria del acopio, el error acumulado a 30 días fue de solo -0.127%, lo que refuerza la confiabilidad de la planeación de abastecimiento. Los resultados evidencian que el modelo descriptivo optimiza la asignación de materia prima, mientras que el modelo predictivo permite pronosticar el ingreso diario de materia prima a la compañía. En conjunto, estos aportes generan beneficios organizacionales al favorecer decisiones más oportunas y alineadas con los objetivos de eficiencia y maximización del rendimiento. En conclusión, el proyecto impulsa la transformación hacia una gestión del acopio más estratégica e inteligente, sentando las bases para un proceso de recepción de leche sustentado en datos, más predecible y orientado a la calidad.

Palabras Clave

Acopio de leche cruda, CRISP-DM, PCA, K-Means, ARIMA, SARIMAX, transformación.

Abstract

This case study addresses the need to transform data-driven decision-making to improve production planning and industrial performance associated with the raw milk collection process in a dairy company. To achieve this, the CRISP-DM methodology was applied through an approach that integrates descriptive and predictive analytics. In the descriptive phase, PCA and K-means techniques were used, identifying three milk segments differentiated by their composition (Premium Milk, Milk-for-Milk, and Standard Milk), which enables the strategic allocation of higher-quality raw material to high-value processes. In the predictive phase, ARIMA and SARIMAX models were evaluated, with parameters (1,0,2)x(1,1,1,30) selected for its superior accuracy, supported by the inclusion of Fat as an exogenous variable and its ability to capture monthly seasonality. Despite the daily volatility of milk collection, the 30-day cumulative error was only –0.127%, strengthening the reliability of supply planning. The results show that the descriptive model improves raw material allocation, while the predictive model allows for highly accurate anticipation of operational needs. Together, these contributions provide organizational benefits by enabling more timely decisions aligned with efficiency and performance goals. In conclusion, the project drives the transition toward a more strategic and intelligent milk collection management approach, laying the groundwork for a data-driven, more predictable, and quality-oriented reception process.

Keywords

Raw milk collection, CRISP-DM, PCA, K-Means, ARIMA, SARIMAX, transition, data-driven decision-making.

1. Introducción

Según un estudio realizado por RADDAR CKG, el 90% de los colombianos consume leche en sus hogares, y el 58% lo hace más de una vez por semana. (Raddar; Asoleche; Ministerio de Agricultura, 2025). Estas cifras evidencian la importancia que tiene el consumo de productos lácteos dentro de los hábitos alimenticios del país. El sector lácteo no solo tiene una fuerte presencia en la canasta básica de los consumidores colombianos, sino que además desempeña un papel estratégico en la economía nacional, tanto por su aporte al PIB agroindustrial como por la generación de empleo directo e indirecto.

Este sector abarca diversas zonas productivas, que van desde la producción primaria de la leche, hasta el procesamiento y distribución de productos derivados lácteos como yogur, queso, crema y leche en polvo. En los últimos años, las innovaciones tecnológicas y las prácticas de mejora continua han permitido a las empresas lácteas optimizar sus procesos, reduciendo costos y mejorando la eficiencia operativa.

Una de las estrategias clave para lograr esta eficiencia consiste en conocer con precisión las características fisicoquímicas de la leche recibida. Esta información permite a las plantas de producción direccionar los distintos lotes hacia procesos específicos según sus propiedades. Por ejemplo, una leche con mayor contenido de grasa puede destinarse a la producción de quesos, mientras que una leche con menor contenido graso es más apta para elaborar leche en polvo o productos bajos en grasa. Esta clasificación no solo mejora el uso de los recursos, sino que también incrementa el rendimiento y reduce los costos asociados a la estandarización del insumo (Fox, & McSweeney, 2015).

En este marco, el presente caso de estudio tiene como fin identificar los comportamientos futuros que nos permita saber cuánta leche es acopiada mensualmente en

la planta de producción de una empresa de lácteos. Así como cuantificar sus principales indicadores de calidad, de tal manera que se puedan generar estrategias desde la producción que puedan hacer más eficientes los costos de transformación de la materia prima.

La metodología propuesta para alcanzar los objetivos expuestos incluye un análisis estadístico descriptivo de los datos disponibles, seguido de una etapa de preparación y validación de la información. Posteriormente, se construirá un modelo de predicción que permita anticipar los comportamientos de acopio y composición de la leche en función de patrones históricos. Esta capacidad predictiva es crucial para direccionar adecuadamente los lotes recibidos hacia productos específicos, evitando el uso ineficiente de leche de menor calidad en productos de mayor valor, y contribuyendo así a la reducción de costos y la mejora de los márgenes operativos (Haug, Høstmark , & Harstad, 2007).

Comenzando con los objetivos general y específicos a los que hará referencia el proyecto, seguido del entendimiento de negocio en el cual se presenta la fundamentación teórica del sector lácteo y la importancia de la caracterización de la leche en los procesos industriales. La comprensión de los datos describe los detalles de la obtención de datos, así como un desarrollo exploratorio expreso mediante estadística descriptiva de las variables con las que se cuenta. Se desarrollan modelos para cada uno de los objetivos que se plantean y así mismo se genera un análisis guiado al entendimiento del negocio. Se cierra mostrando las conclusiones y la bibliografía utilizada.

2. Objetivo

2.1. Objetivo General

Optimizar la gestión productiva de una compañía láctea a través de un modelo analítico que combine la segmentación por calidad de la leche y la predicción de volúmenes de acopio para potenciar la toma de decisiones estratégicas.

2.2. Objetivos Específicos

- Implementar la metodología de datos CRISP-DM como marco estructurado para el desarrollo del proyecto de segmentación por calidad y la predicción de volúmenes de acopio, abarcando desde el entendimiento del negocio y los datos, hasta la preparación, modelado, evaluación y despliegue de los resultados analíticos.
- Aplicar técnicas de clusterización no supervisada, apoyadas en reducción dimensional mediante PCA, para identificar segmentos de la leche acopiada en grupos homogéneos según su perfil composicional y sus atributos de calidad con el fin de hacer más eficiente su distribución en las líneas de producción.
- Sugerir modelos de predicción de volúmenes de acopio utilizando técnicas de series de tiempo con el fin de anticipar patrones de comportamiento y variabilidad estacional en las entregas de leche mejorando la planificación operativa del negocio.

3. Comprensión del negocio

3.1. Generalidades de la leche

La leche es un alimento integral en la nutrición humana, compuesta por aproximadamente un 87 % de agua y un 13 % de sólidos, incluyendo grasas (~3.5–4 %), proteínas (~3–3.5 %), lactosa (~4.8 %), minerales (~0.7 %), y vitaminas (A, D, E, K, B2, B12) (FAO, 2025). Las proteínas de alto valor biológico (80 % caseínas, 20 % proteínas de suero) y los lípidos, principalmente triacilglicéridos, le confieren una fuente importante de energía y nutrientes esenciales (Kourkouta et al., 2021). Esta composición convierte a la leche en una materia prima extremadamente versátil para la industria alimentaria, permitiendo su transformación en leche líquida pasteurizada o UHT, quesos, yogures, mantequilla y leche en polvo. Sin embargo, por su naturaleza biológica, perecedera y variable, la conservación y procesamiento industrial dependen críticamente de sus características fisicoquímicas y microbiológicas, las cuales están condicionadas por factores como el clima, la alimentación del ganado, prácticas de higiene y la cadena de frío (Kourkouta et al., 2021).

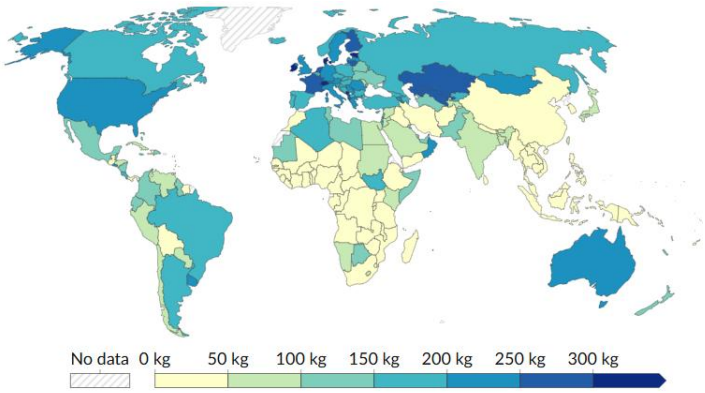
3.2. Consumo per Cápita de Leche: Contexto Global y Desafíos para Colombia

A nivel global, el consumo promedio de leche y productos lácteos se sitúa en 119 litros por persona al año (OCLA, 2023), lo que representa aproximadamente el 60 % del consumo recomendado por la FAO-OMS, que establece debe estar en el orden de los 180 l/año (OCLA, 2016; Tapia, 2020). Tal como se muestra en la Figura 1., existen marcadas diferencias entre regiones: en los países desarrollados el consumo supera los

240 l/año, mientras que en países en desarrollo apenas alcanza unos 80 l/año (Our World in Data, 2022).

Figura 1.

Consumo per cápita de leche (en kg por persona/año) por país en 2022.

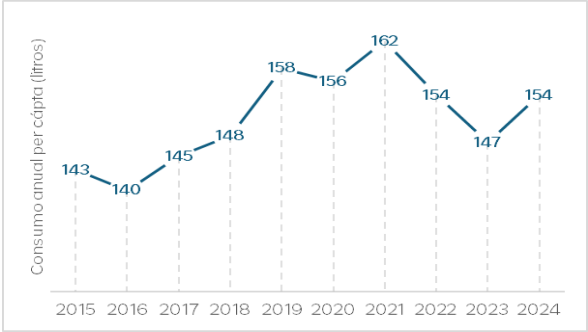


Fuente: Tomada de *Per capita milk consumption [Mapa interactivo]*. Por Our World in Data, 2022.

En Colombia, el consumo per cápita estimado es de aproximadamente 154 litros anuales, por debajo de las recomendaciones nutricionales y del promedio de países de la OCDE, pero por encima de muchas otras economías emergentes (Fedegan, 2024). Esta cifra refleja tanto el crecimiento del acceso a productos lácteos como los retos persistentes en términos de equidad, hábitos alimentarios y capacidad adquisitiva en el país.

Figura 2.

Consumo aparente per cápita anual Leche en Colombia 2015 - 2024



Fuente. Adaptado de *Consumo aparente per Cápita anual de Leche*, por Fedegan, 2024.

3.3. Panorama General del Sector Lácteo en Colombia

El sector lácteo colombiano representa uno de los pilares fundamentales de la economía agroindustrial del país, con más de 7.000 millones de litros de leche producidos en 2023, presentando un crecimiento del 8% en la última década. (Our World in Data, 2023). Se caracteriza por una alta fragmentación en la producción primaria, con miles de pequeños y medianos productores distribuidos en zonas rurales de clima templado y frío, como Cundinamarca, Antioquia, Boyacá y Nariño.

El abastecimiento presenta una marcada estacionalidad, con caídas productivas durante las temporadas secas y picos en épocas de lluvias, lo que afecta tanto la cantidad como la calidad del insumo.

3.4. Modelo Operativo de la Compañía

Una de las funciones principal del equipo de manufactura en una empresa del sector alimentario radica en asegurar un margen operativo que permita a la organización mantener flexibilidad a lo largo de la cadena de valor. Esta flexibilidad resulta fundamental para la consecución del objetivo de EBITDA (Earnings Before Interest, Taxes, Depreciation, and

Amortization), al facilitar la adaptación a las condiciones cambiantes del mercado y la implementación de estrategias competitivas (Christopher & Holweg, 2011).

La compañía objeto del estudio posee una red nacional de centros de acopio, plantas de transformación ubicadas estratégicamente, y una capacidad instalada que le permite competir por ser el líder en el mercado nacional en categorías clave como leche líquida, quesos y derivados fermentados.

El modelo operativo se sustenta en la recepción diaria de leche cruda proveniente de miles de proveedores, a través de rutas de recolección directa y de intermediarios regionales. Una vez recepcionada, la leche es sometida a análisis fisicoquímicos y microbiológicos para verificar el cumplimiento de los requisitos de ingreso como materia prima a la planta. Estos controles se realizan mediante equipos de detección automática que miden parámetros como grasa, proteína, sólidos no grasos (SNG), sólidos totales (ST), densidad, entre otros, los cuales determinan su aptitud para los distintos procesos industriales.

Actualmente, la asignación de la leche a las líneas de producción, como leche UHT, yogur o quesos, se gestiona bajo el criterio PEPS (Primero en Entrar, Primero en Salir), considerando exclusivamente la hora de llegada una vez aprobada por calidad. Sin embargo, este enfoque limita el aprovechamiento de la materia prima, ya que no considera las diferencias en composición entre los distintos lotes, como el contenido de grasa, proteína o sólidos no grasos. En la práctica, esto conduce a la mezcla de leches con estándares altos y bajos, generando una composición homogénea con valores promedio, lo que reduce el potencial de aprovechamiento y la rentabilización en productos específicos

3.5. Variables Críticas en el Proceso de Acopio

El volumen y calidad de la leche dependen de una combinación de factores:

- **Estacionales:** El ciclo climático (invierno/verano) afecta la producción por disponibilidad de pasto y agua.
- **Zonales:** Las condiciones agroclimáticas y la tecnología de producción varían significativamente entre regiones.
- **Productor-dependientes:** Prácticas de alimentación, manejo sanitario y conservación influyen en el contenido de grasa, proteína, acidez y presencia de residuos.
- **Económicos:** Los precios pagados al productor, el costo de insumos y los incentivos de calidad impactan el suministro.

La variabilidad en la composición de la leche implica que no toda materia prima tiene el mismo potencial de rendimiento según el tipo de producto a elaborar, afectando directamente su rentabilidad. Los indicadores de calidad de la leche, como sólidos totales, porcentaje de grasa y porcentaje de proteína, son determinantes para la asignación del insumo y el rendimiento industrial. Por ejemplo, según la experiencia de la compañía, un mayor contenido de proteína se traduce en un incremento en el rendimiento quesero, permitiendo producir una mayor cantidad de queso con el mismo volumen de leche. De manera similar, las leches con mayor contenido graso posibilitan una mayor obtención de crema, la cual puede destinarse a la elaboración de productos de alto valor comercial, como cremas pasteleras o mantequillas.

3.6. Gestión Actual de la Planificación

La planificación del ingreso de leche al proceso industrial se realiza actualmente con base en promedios históricos, reportes semanales y estimaciones empíricas del volumen esperado. Esta aproximación presenta diversas limitaciones, entre las que se destacan:

- Escasa anticipación para ajustar los planes de producción en función de la disponibilidad real de leche que abastece a la planta.
- Necesidad permanente de realizar ajustes reactivos, lo que afecta negativamente la eficiencia operativa.
- Activación tardía de negociaciones externas por parte de áreas como Compras y Planeación, dificultando la gestión oportuna de excedentes o déficits de leche con proveedores o clientes.

La anticipación de los niveles de ingreso de leche abre la oportunidad de mitigar las limitaciones anteriormente descritas. En este sentido, la incorporación de herramientas predictivas permitiría adoptar un enfoque preventivo, facilitando la toma de decisiones oportunas y mejorando la gestión operativa de la planta.

3.7. Justificación del proyecto

Este proyecto busca dotar a la compañía de un modelo predictivo de acopio y segmentación de leches que permita anticipar con mayor precisión el volumen mensual esperado y sus principales indicadores de calidad. Esto permitirá:

- Optimizar la planificación industrial con criterios de calidad anticipada.
- Maximizar el rendimiento técnico y económico de cada litro procesado.
- Reducir costos ocultos asociados a decisiones reactivas.
- Dar un paso hacia la transformación digital al inicio de la cadena productiva en la recepción de leche.

En síntesis, se trata de aprovechar el valor de los datos históricos y contextuales para lograr una gestión más inteligente y rentable del principal insumo de la compañía: la leche cruda.

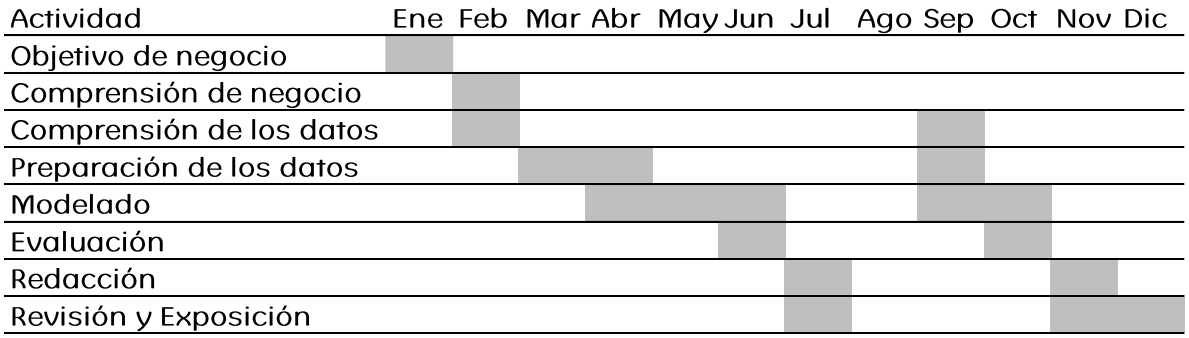
3.8. Plan de Proyecto

El proyecto se desarrolló de acuerdo con el cronograma presentado en el diagrama Gantt de la Figura 3., éste se encuentra organizado por meses y abarca el período comprendido entre marzo y noviembre de 2025. El análisis del cronograma evidencia un comportamiento cíclico del proceso. Por ejemplo, durante el mes de febrero, se ejecuta la etapa de comprensión y preparación de datos orientada al desarrollo del modelo de clusterización. Posteriormente, en septiembre, se retoma nuevamente la fase de comprensión de datos, en esta ocasión enfocada en el diseño del modelo de series de tiempo.

El proyecto concluye en el mes de noviembre, momento en el cual se consolidan y cumplen tanto el objetivo operativo como el objetivo estratégico, cuyos resultados se presentan en el apartado siguiente.

Figura 3.

Cronograma de actividades para el año 2025.



Fuente: Elaboración propia.

4. Comprensión de los datos

4.1. Recopilación de Datos Iniciales

Se cuenta con dos bases de datos tomadas del ERP que trabaja la compañía. Las bases, descargadas en formato Excel, contienen información producida por la organización durante el año 2022 a 2024, la cual consolida la fecha, información proveniente del vehículo que recoge la materia prima y la entrega a la compañía, así como los factores de calidad relevantes (% de grasa, % de proteína, % de solidos no grasos, % de solidos totales). Dependiendo de la variable varía su formato, existen variables numéricas, categóricas y de fecha. Esta base se actualiza a diario, sin embargo, para fines académicos del presente proyecto, se toman datos históricos.

Para la base de datos existen variables con información sensible para la empresa y otros poco relevantes para el estudio. Por lo anterior no serán tenidas en cuenta, por ejemplo, los lotes de inspección propios de la muestra, o información que referencie nombres o códigos internos de la compañía.

4.2. Descripción de los Datos

4.2.1. Base de datos de Calidad de Leche

Esta base tiene por objetivo el registro de datos de calidad clave para cada uno de los lotes recepcionados, se dispone de un total de 41.969 registros, con 12 variables seleccionadas que serán utilizadas en el desarrollo del proyecto.

Tabla 1.

Información 2022 - 2024 de calidad de leche cruda acopiada.

Campo	Variable	Tipo de Dato	Descripción
1	Centro	int64	Número de planta de recepción
2	Inicio de Inspección	datetime64[ns]	Fecha de análisis de la entrega de leche
3	Lote	object	Código de ruta de recolección
4	Material	int64	ID del productor o tipo de leche
5	Texto Breve de Material	object	Descripción del número de material
6	ACIDEZ04	float64	Acidez Titulable
7	pH	float64	pH de la muestra
8	DENSID01	float64	Densidad
9	GRASAEQU	float64	Materia grasa (%)
10	PROTE08	float64	Proteína (%)
11	SOLID06	float64	Sólidos no grasos (%)
12	SOLID07	float64	Sólidos totales (%)

Fuente: Elaboración propia.

4.2.2. Base de datos de Volumen Acopiado

Esta base tiene por objetivo registrando el volumen entregado por cada lote de leche cruda acopiada. Cuenta con un total de 44.413 registros y 14 variables. Sin embargo, por tratarse de información sensible, a continuación, se presentan únicamente las variables seleccionadas para el desarrollo del proyecto.

Tabla 2.

Información 2022 - 2024 de volumen de leche cruda acopiada.

Campo	Variable	Tipo de Dato	Descripción
1	Fecha	datetime64[ns]	Fecha de la entrega de leche
3	Lote	object	Código de ruta de recolección
5	Vol. Litros	int64	Volumen descargado de la ruta
13	Centro	int64	Número de planta de recepción

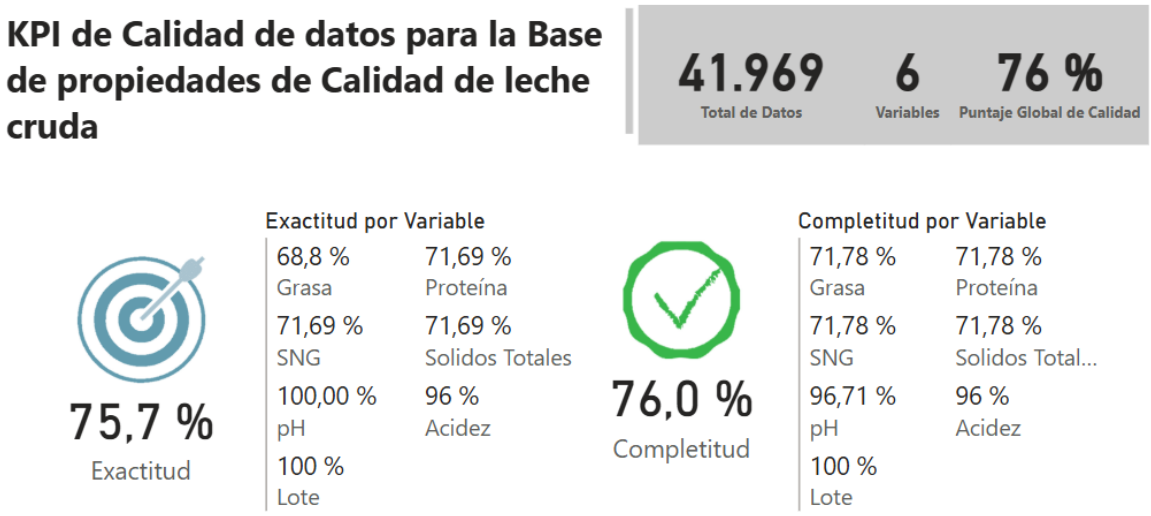
Fuente: Elaboración propia.

4.3. Calidad de los Datos

La Figura 4 presenta el análisis preliminar de calidad de datos diseñado con la herramienta PowerBI. Los principales indicadores clave de desempeño (KPI) asociados a la calidad de los datos de la base correspondiente a las propiedades de calidad fisicoquímica de la leche cruda. En total, se analizan 41.969 registros, distribuidos en 6 variables, alcanzando un puntaje global de calidad del 76%. Este resultado refleja un desempeño general aceptable; sin embargo, evidencia oportunidades de mejora, particularmente en el atributo de exactitud.

Figura 4.

Análisis de Calidad de Datos para la Base de datos de propiedades de calidad de leches acopiadas.



Fuente: Elaboración propia.

La exactitud, entendida como la fidelidad de los datos frente a su valor real o esperado, presenta un valor promedio de 75,7%, mientras que la completitud, que mide la proporción de datos disponibles respecto al total esperado, alcanza un 76%, situándose

ligeramente por encima. Ambos indicadores muestran comportamientos similares a nivel agregado, pero con diferencias relevantes al analizarse por variable.

Al desagregar los KPI por variable, se observa una variabilidad significativa en los niveles de exactitud y completitud. Las variables Lote, pH y Acidez presentan valores superiores al 96%, destacándose frente al resto de las variables, cuyos indicadores se sitúan en rangos entre 68,8 % y 71,78%. Esta diferencia se explica por la estrategia operativa de muestreo aplicada en planta: debido a restricciones asociadas al costo de las pruebas y a la disponibilidad limitada de recursos analíticos, no se evalúan de manera rutinaria todas las propiedades de calidad en el 100% de las muestras. En su lugar, se prioriza la medición de pH y acidez, al ser indicadores críticos para viabilizar el recibo de la materia prima. Este enfoque selectivo, si bien es operativo y técnicamente justificado, genera vacíos de información en otras variables, lo que impacta negativamente la calidad global de la base de datos.

Adicionalmente, es importante aclarar que el atributo de integridad no fue considerado dentro de este análisis con el fin de evitar falsos positivos. En este contexto, la presencia de datos vacíos no implica una pérdida de integridad del dato, sino que responde a la no ejecución de determinadas pruebas, de acuerdo con el esquema de muestreo definido. Por tanto, evaluar integridad podría inducir a conclusiones erróneas sobre la calidad real de la información.

Finalmente, la consistencia de los datos se considera garantizada con un 100% de confianza, dado que el ERP utilizado no permite la existencia de más de un registro con el mismo lote y fecha para una entrega de leche en planta, asegurando así la ausencia de contradicciones estructurales dentro de la base de datos. Para el desarrollo del proyecto, se

llevarán a cabo proceso de limpieza y transformación, los cuales serán explicados en sesiones posteriores.

4.4. Exploración de los Datos

A continuación, se presenta la descripción de los datos relevantes para el proyecto, considerando el entendimiento del negocio y la necesidad de realizar un análisis tanto general como segmentado estratégicamente por meses. Esta decisión responde a la influencia de la estacionalidad previamente mencionada, la cual impacta directamente en la calidad y volumen de la leche.

En la presente sección se usará el software R Studio como medio de análisis estadístico mediante el uso de librerías tales como *readxl, psych, ggplot2*.
Para facilitar la interpretación, el análisis descriptivo se organizará de manera detallada para cada una de las variables que aporten información significativa y de valor para los objetivos del proyecto. Se excluirán aquellas variables que resulten redundantes o poco relevantes, siempre garantizando la confidencialidad y seguridad de los datos.

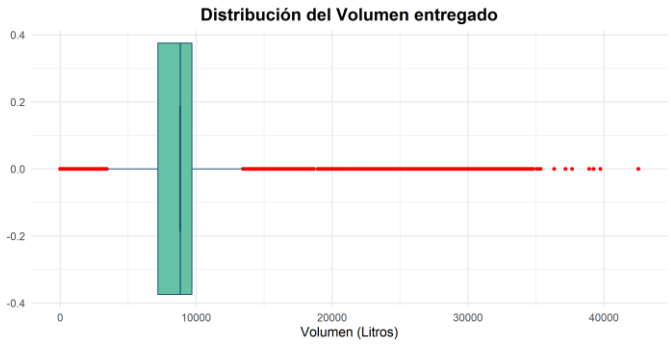
4.4.1. Volumen

Al realizar un análisis estadístico de las entregas realizadas individualmente, encontramos una generalidad que se encuentra en función a la capacidad del vehículo que hace la entrega de leche a planta. Tal como se expone en el gráfico de caja en la Figura 5, la distribución del volumen de leche acopiada por entrega cuenta con una marcada asimetría positiva. La mayoría de los registros se concentran en un rango intercuartílico que oscila entre los 7.100 a 9.700 litros, siendo el volumen típico de entrega. No obstante, se

presentan numerosos valores superiores a éste rango, al tratarse de movimientos de leche hechos en contenedores de mayor volumen, que pueden llegar a ser tractomulas, cuya capacidad puede oscilar los 30.000 litros en su contenedor.

Figura 5.

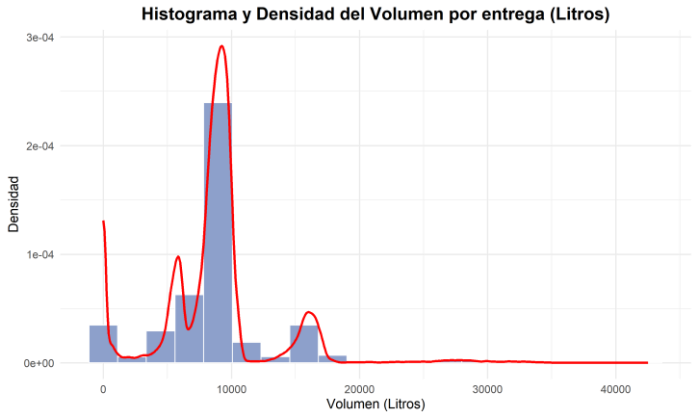
Boxplot - Distribución del volumen por lote entregado



Fuente: Elaboración propia.

Figura 6.

Histograma - Distribución del volumen por lote entregado



Fuente: Elaboración propia.

Para complementar el análisis anterior, la Figura 6 presenta un histograma que permite observar con mayor detalle la distribución del volumen entregado por lote de leche

cruda. Esta distribución muestra una asimetría positiva, con un pico principal entre los 9.000 y 10.000 litros, lo que indica que la mayoría de las entregas se concentran en ese rango. Además, la curva de densidad sugiere una distribución multimodal, lo cual valida la existencia de distintos tipos de capacidad en los contenedores utilizados (pequeños, medianos y grandes). Este patrón refuerza la necesidad de considerar una normalización de la muestra para los análisis posteriores, a fin de evitar sesgos en la interpretación de los datos.

Asimismo, se identifican registros con valores de volumen igual a cero, los cuales deben ser excluidos del estudio, ya que se presume que corresponden a errores de digitación o fallos en el cargue de información en el ERP. Esta decisión se sustenta en el conocimiento del negocio, donde no resulta válido registrar un descargue equivalente a cero litros, por lo que estos datos no aportan valor y pueden distorsionar los resultados del modelo.

4.4.2. Variables de Calidad de leche

En la Tabla 3 se presenta un resumen de las variables de Calidad de leche con las que se cuenta en la base de datos. En aspectos generales, las variables presentan una distribución relativamente estable para pH, acidez, grasa, proteína, SNG (Sólidos no grasos) y ST (Sólidos totales), con medias y medianas muy cercanas entre sí, lo que indica simetría en los datos.

Tabla 3.

Tabla estadística descriptiva para variables de calidad de leche.

Variable	pH	Acidez	Grasa	Proteína	SNG	ST
Min	5.35	8	0	0	0	0
1er Cuartil	6.74	14.2	3.60	3.10	8.34	12.03
Mediana	6.76	14.4	3.76	3.16	8.43	12.22
Media	6.756	14.4	3.83	3.158	8.426	12.26
3er Cuartil	6.78	14.6	3.98	3.22	8.53	12.44
Max	6.80	36.4	14.58	3.68	9.19	41.19
Curtosis	130	836	33	255	475	356
NA's	1379	1488	11844	11843	11843	11842

Fuente: Elaboración propia.

La variable *acidez* presenta una notable dispersión, alcanzando valores de hasta 36,4, lo cual podría deberse a errores de medición o al análisis tardío de ciertas muestras, situación que puede comprometer la precisión de los resultados.

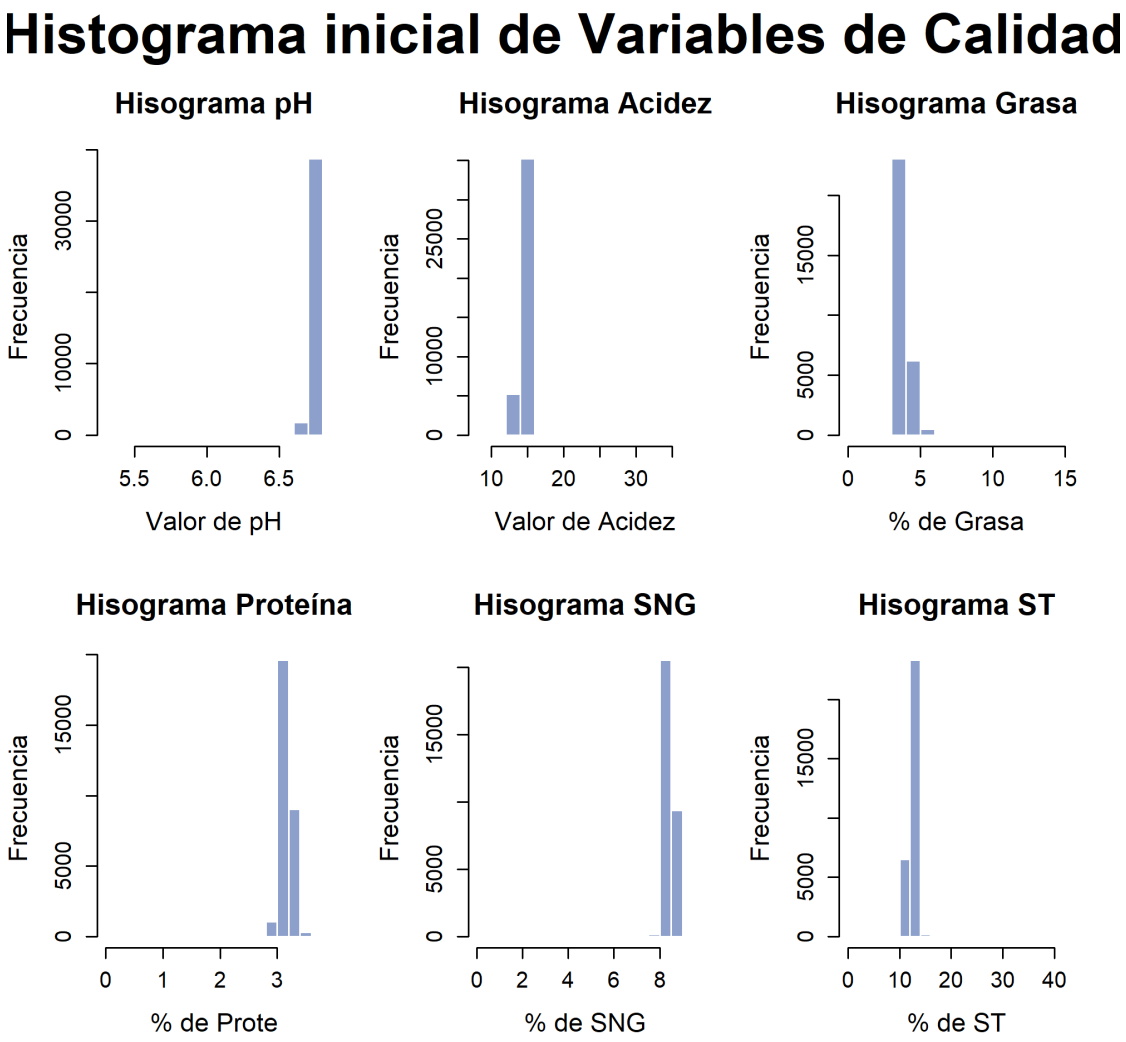
Por otro lado, al analizar la curtosis de los datos, se evidencia en todos los casos un comportamiento leptocúrtico (ver Figura 7), caracterizado por picos marcadamente pronunciados y colas pesadas. Esta distribución indica la presencia significativa de valores atípicos, lo que refuerza la importancia de realizar una limpieza rigurosa del conjunto de datos.

Finalmente, y en línea con lo planteado en el numeral 4.3, se identifican valores mínimos de cero en las variables *grasa*, *proteína*, *SNG* y *ST*, situación que no es fisiológicamente posible y que se explica por la práctica operativa de realizar muestreos parciales, centrados únicamente en *pH* y *acidez*. Estas mismas variables presentan un alto número de datos faltantes (alrededor de 11.844 registros por variable), lo cual afecta de

forma considerable la calidad del conjunto de datos y deberá ser abordado en el proceso de limpieza y análisis posterior.

Figura 7.

Histograma de datos de Calidad previos a tratamiento de datos.



Fuente: Elaboración propia.

5. Preparación de los datos

5.1. Selección de los datos

Como resultado del análisis descriptivo, el entendimiento de los datos y del negocio, se identifican las siguientes variables para adaptarlos a las técnicas de Data Mining:

- Fecha (Enero 2022 a Diciembre 2024)
- Lote (83 niveles, R0001, R0002, R0003,…,R0110)
- Vol. Litros (Volumen descargado de la ruta)
- ACIDEZ04 (Grados de acidez titulable del lote a descargar)
- PH (pH de la muestra de la ruta a descargar)
- GRASAEQU (% Grasa de la muestra de la ruta a descargar)
- PROTE08 (% Proteína de la muestra de la ruta a descargar)
- SOLID06 (% Sólidos no grasos de la muestra de la ruta a descargar)
- SOLID07 (% Solidos Totales de la muestra de la ruta a descargar)

5.2. Construcción de nuevos datos

Para el desarrollo de los modelos, cada base de datos recibirá un tratamiento específico. Inicialmente, se utilizará la información de calidad para generar, mediante clusterización, una nueva variable que clasifique las entregas de leche. Posteriormente, este resultado se integrará con la base de datos restante para consolidar la información final necesaria para el modelo predictivo de series de tiempo.

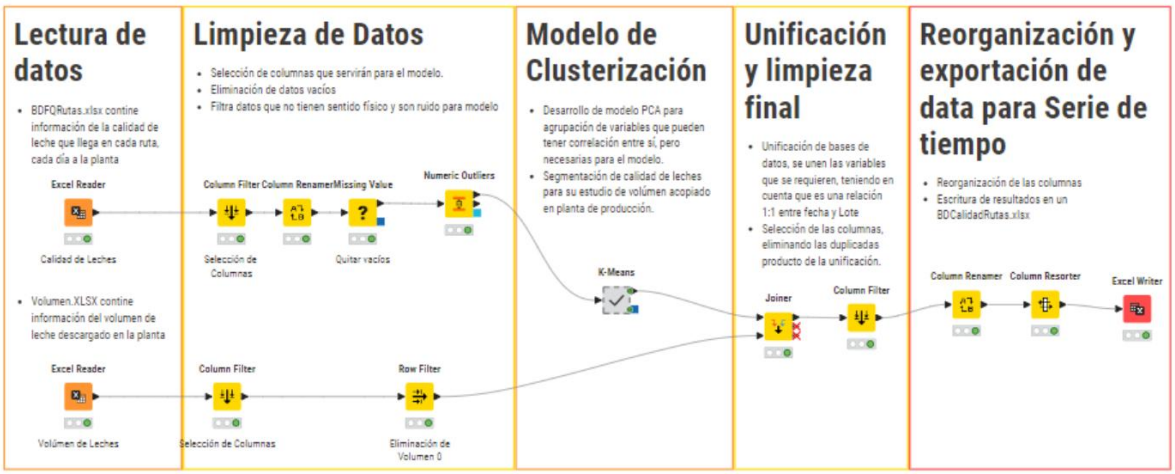
5.3. Extracción y transformación de datos

Para llevar a cabo el proceso de preparación de los datos se usará el Software KNIME Versión 5.4.4. El objetivo de la preparación de los datos se basará de la selección de columnas y filas que no tengan sentido físico desde el entendimiento de negocio, así como

de la unificación de variables que se encuentran distribuidas en dos bases de datos. El proceso general se observa en la Figura 8 y se explica a continuación.

Figura 8.

Lectura, limpieza y unificación de datos.



Fuente: Elaboración propia.

5.3.1. Lectura de Datos

Se toman dos bases de datos en las cuales se encuentran las variables seleccionadas en la sección 5.1. Inicialmente, la base *BDFQRutas.xlsx (Base 1 en lo que sigue)* que recupera información de calidad de las leches, cuenta con un total de 12 variables y 41.969 filas. Por otra parte, la base *Volumen.xlsx (Base 2 en lo que sigue)* contiene datos referentes al volumen de acopio, con un total de 14 variables y 44.413 filas.

5.3.2. Limpieza de Datos

Inicialmente se seleccionan las variables que se requieren para cada una de las bases de datos, para la Base 1 se dejan 8 variables, reduciendo 4. Por su parte, la Base 2 queda con 3 variables, de 14 totales.

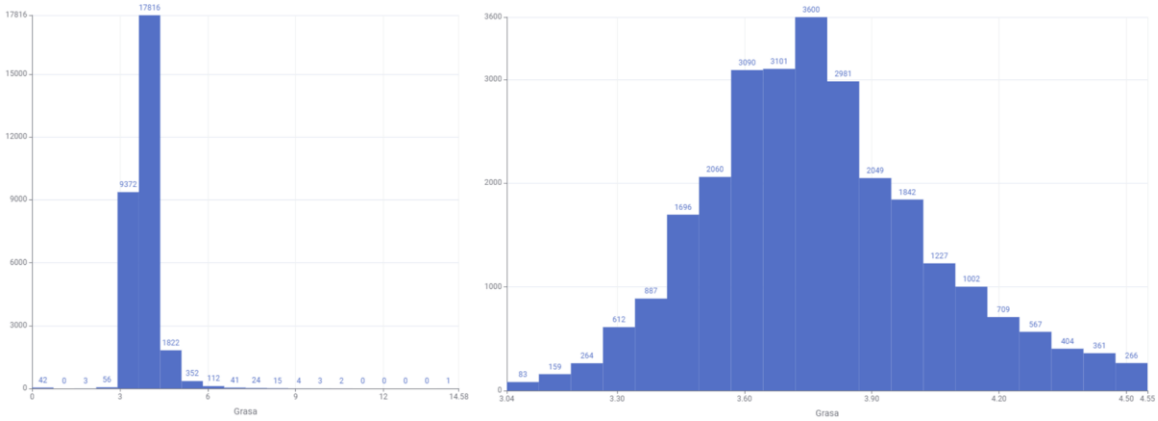
A continuación, se lleva a cabo una primera eliminación de información faltante, mediante un nodo ‘Missing Value’ con el que buscamos reducir toda fila que pueda tener datos vacíos en la Base 1, teniendo en cuenta que es información faltante por la razón de negocio expuesta en la sección de calidad de datos (ver sección 4.3.).

Luego se procede a eliminar registros que carecen de sentido físico o de negocio, como los siguientes casos:

- En el caso de los datos de la Base 1 ‘Calidad de leche’, se realiza una limpieza de valores atípicos en cada una de las variables. Para tomar un ejemplo, la variable Grasa, basada en parámetros fisiológicos y bibliografía especializada. Según la FAO (2025), el contenido de grasa en leche de vaca suele oscilar entre el 3 % y el 4 %.

Figura 9.

Comparación de histogramas luego de la eliminación de Outliers para la variable Grasa



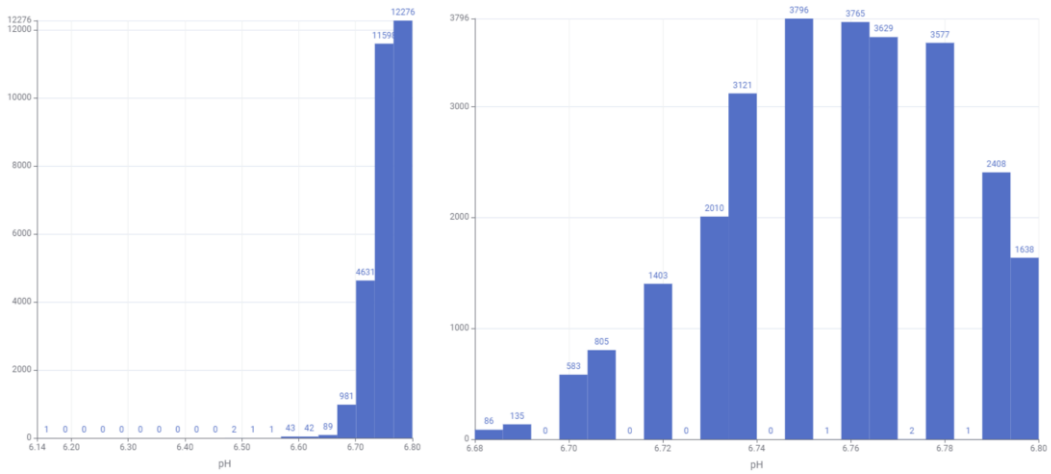
Fuente: Elaboración propia.

- Finalmente, observando el comportamiento de frecuencias de la variable pH (ver Figura 10) se observan dos datos, correspondiente al 0.005% de los datos, los cuales pueden tratarse de errores de cargue de información y que al retirarlos de la base reducen la curtosis de la variable de pH de 130 a 0.06, y de la Acidez titulable de

836 a 3.11, siendo estadísticamente mucho más valioso, a cambio de la perdida insignificante de información.

Figura 10.

Histograma comparativo de pH tras la eliminación de valores atípicos



Fuente: Elaboración propia.

- Para la Base 2, los registros con volumen igual a cero, los cuales corresponden a rutas que figuran en el sistema pero que, en la práctica, no realizaron ninguna descarga. Este error sistemático puede introducir ruido y afectar negativamente el desempeño de los modelos. Se reducen 2.935 datos, el 6.6% del total de registros.

Como resultado del proceso de limpieza, la Base 1 contiene ahora el 64% de la información inicial con 26.960 registros, mientras que la Base 2 mantiene el 93% de la información inicial con 41.478 registros.

5.3.3. Verificación Estadística

En la Tabla 4. se exponen los cambios tras aplicar el proceso de limpieza de datos, se observan mejoras significativas en la calidad y consistencia de las variables analizadas. En primer lugar, se eliminaron por completo los valores vacíos en todas las variables,

garantizando una base completa para el análisis. Además, se redujeron considerablemente los valores extremos (outliers), lo que se refleja en una disminución marcada de la asimetría (Sk.) y la curtosis (Kurt.) en todas las variables. Por ejemplo, la acidez pasó de una asimetría de 17 y una curtosis de 837 a valores mucho más estables de 0.4 y -0.43, respectivamente. Lo mismo ocurre con pH, cuya asimetría extrema inicial de -3.9 se redujo a -0.37, y su curtosis bajó de 131 a -0.36. También se observa una contracción general en la varianza y desviación estándar, lo que indica una mayor homogeneidad y menor dispersión de los datos. En resumen, la limpieza de datos logró corregir distorsiones importantes en la distribución, estabilizando los estadísticos descriptivos y mejorando sustancialmente la calidad del conjunto de datos para su posterior análisis.

Tabla 4.

Estadística descriptiva comparativa entre variables antes y después de limpieza de datos.

Estadística Descriptiva de variables de Calidad de leches inicial										
Variable	Vacíos	Mínimo	Máximo	25%	Media	75%	D. Est.	Var.	Sk.	Kurt.
Acidez	1488	8	36.4	14.2	14.4	14.6	0.37	0.135	17	837
pH	1379	5.35	6.8	6.74	6.76	6.78	0.03	0.001	-3.9	131
Grasa	11844	0	14.58	3.6	3.76	3.98	0.48	0.229	2.5	34
Proteína	11843	0	3.68	3.1	3.16	3.22	0.15	0.023	-12	255
SNG	11843	0	9.19	8.34	8.43	8.53	0.34	0.118	-20	476
ST	11842	0	41.19	12.03	12.22	12.44	0.67	0.444	-2.7	357

Estadística Descriptiva de variables de Calidad de leches luego de Limpieza de datos										
Variable	Vacíos	Mínimo	Máximo	25%	Media	75%	D. Est.	Var.	Sk.	Kurt.
Acidez	0	14	15.2	14.2	14.4	14.6	0.26	0.069	0.4	-0.43
pH	0	6.68	6.8	6.74	6.76	6.78	0.03	0.001	-0.37	-0.36
Grasa	0	3.04	4.55	3.59	3.75	3.93	0.26	0.072	0.4	0.097
Proteína	0	2.92	3.4	3.1	3.16	3.22	0.08	0.007	0.15	-0.24
SNG	0	8.06	8.81	8.35	8.44	8.53	0.14	0.018	0.08	-0.21
ST	0	11.42	13.05	12.03	12.2	12.4	0.28	0.076	0.25	-0.07

Fuente: Elaboración propia.

5.3.4. Unificación y Limpieza Final

En esta sección se lleva a cabo la unificación de las dos bases de datos partiendo del criterio de unicidad entre las columnas de lote y fecha, la intersección de estas dos variables genera la unión de la información y generando una única tabla con las variables que se requieren.

En esta sección la información que se unifica para ambas bases de datos corresponde a 27.891 datos, un 66% para la Base 1 y un 63% para la Base 2.

5.3.5. Reorganización y exportación de data

En esta sección se adecúan los nombres de las columnas, se reorganiza la información por variables y se dejan disponibles en un formato Excel para su almacenamiento.

5.3.6. Análisis del costo – beneficio de la limpieza de datos

En la Tabla 5 se muestra cómo a medida que avanza el tratamiento de datos, trae consigo un costo asociado a la pérdida de información.

Tabla 5.

Costo - Beneficio de la preparación de los datos

Base/Proceso	**Lectura**		**No Vacíos**		**Precisos**		**Unificación**		**Salida**	
Base 1	41969	100%	29655	71%	27898	66%	27891	66%	27891	66%
Base 2	44413	100%	44413	100%	41478	93%	41478	93%	41478	93%

Fuente: Elaboración propia.

El análisis conjunto de las dos bases de datos permite evidenciar con claridad el equilibrio entre costo y beneficio en el proceso de limpieza y consolidación de datos. Aunque ambos conjuntos contaban inicialmente con un volumen considerable de registros (Base 1: 41.969 y Base 2: 44.413), la aplicación de filtros de calidad, tales como la eliminación de datos vacíos, la revisión de coherencia respecto al negocio y la unificación

de bases provocó una reducción significativa en la cantidad de datos utilizables. En la Base 1, únicamente el 66 % de los registros originales cumplió con los criterios establecidos para su uso en el análisis. Por su parte, la Base 2, a pesar de no presentar vacíos y partir con el 100 % de completitud, redujo su volumen útil al 93 % tras el proceso de integración con la Base 1.

Este proceso de depuración representa una disminución controlada en la cantidad de datos, pero con una mejora sustancial en su calidad estadística. Tal como se evidenció en la sección anterior, dicha mejora se refleja en la reducción de métricas como la asimetría, la curtosis y la varianza. En términos prácticos, se asume una pérdida del 34 % al 37 % de los registros, lo cual se ve ampliamente compensado por el aumento en precisión, confiabilidad y solidez del análisis. Esta limpieza de datos permite el desarrollo de modelos predictivos y descriptivos más robustos y representativos, al eliminar el impacto negativo de valores nulos, inexactos o inconsistentes.

Tras completar la limpieza de información, se identificó una reducción en el periodo temporal disponible para el análisis. Inicialmente, se contaba con registros desde enero de 2022 hasta diciembre de 2024. Sin embargo, tras la depuración, se observó que los datos completos comienzan a partir de abril de 2022. Esta reducción en el rango de fechas puede atribuirse a un proceso de migración tecnológica en los sistemas de registro de la compañía, lo cual habría afectado la captura o conservación de datos en los primeros meses del año 2022.

En conclusión, aunque se sacrifica una parte del volumen inicial de información, el proceso de depuración aporta un valor analítico significativo, asegurando que los datos que se incorporan al análisis final sean pertinentes, fiables y de alta calidad.

6. Modelado - Clusterización

Con el objetivo de identificar patrones de calidad en la leche acopiada y facilitar su clasificación operativa, se aplicó una técnica de clusterización no supervisada mediante el algoritmo K-means, en combinación con un Análisis de Componentes Principales (PCA) para la reducción de dimensionalidad. El algoritmo K-means fue elegido por su eficiencia computacional, facilidad de implementación y capacidad para segmentar grandes volúmenes de datos en grupos homogéneos, en función de similitudes internas. En este caso, permitió agrupar los registros de entrega de leche según sus características composicionales (grasa, proteína, sólidos no grasos y sólidos totales), maximizando la cohesión dentro de los grupos y su separación entre ellos.

6.1. Análisis de correlación entre variables

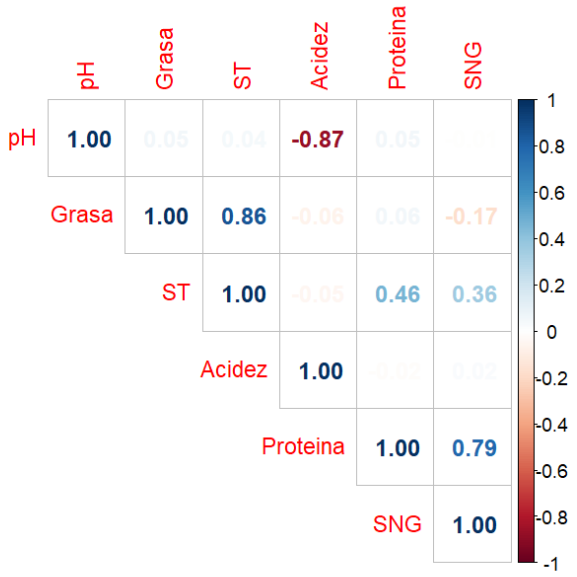
Para evaluar la relación entre las variables fisicoquímicas de la leche acopiada, se realizó un análisis de correlación que se muestra en la Figura 11, acompañado del cálculo de p-valores para determinar la significancia estadística de los coeficientes obtenidos. Los resultados mostraron correlaciones fuertes y significativas entre algunas variables clave:

- La grasa presentó una alta correlación positiva con los sólidos totales ($r = 0.859$, $p < 0.001$), algo que tiene sentido físico, ya que si observamos la definición matemática observamos que los sólidos totales son los sólidos no grasos sumados con la grasa. Lo que indica una dependencia directa.
- La proteína se correlacionó significativamente con los sólidos no grasos ($r = 0.786$, $p < 0.001$). Lo anterior entendiendo que los sólidos no grasos en la leche están compuestos por proteína, lactosa, vitaminas y minerales.

- Las variables como el pH y la acidez mostraron una correlación negativa fuerte (r = -0.87, p < 0.001), coherente con el comportamiento químico esperado, toda vez que el pH es una medida directa de qué tan ácido es un fluido, de tal modo que a una mayor acidez se obtiene un menor valor de pH. Sin embargo, no guardan una relación relevante con las variables composicionales de calidad.

Figura 11.

Análisis de correlaciones de las variables de calidad.



Fuente: Elaboración propia.

A partir de los hallazgos previos, se identifica la necesidad de reducir la dimensionalidad del conjunto de datos a través de dos enfoques complementarios. En primer lugar, se decide eliminar las variables Acidez y pH, debido a su alta correlación entre si y además una baja significancia en la relación con las variables restantes. En segundo lugar, se aplicará una técnica de reducción dimensional mediante Análisis de Componentes Principales (PCA), con el objetivo de mitigar problemas de multicolinealidad

entre variables y, al mismo tiempo, mejorar la robustez y eficacia del modelo de clusterización.

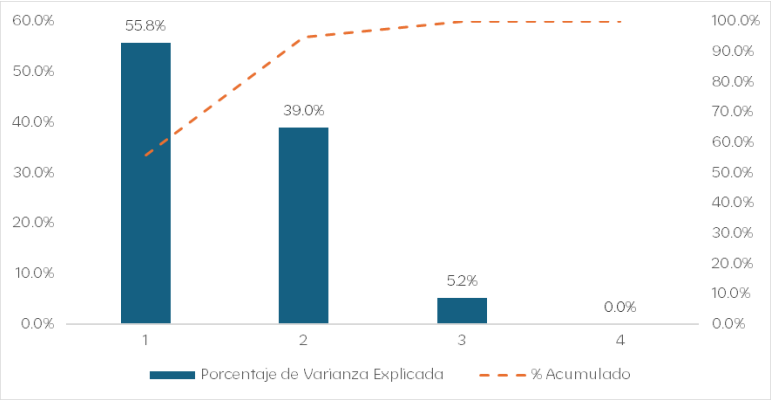
6.2. Análisis por componente (PCA)

Con el objetivo de mitigar los efectos de la multicolinealidad entre variables, se llevó a cabo un análisis de componentes principales (PCA). Este procedimiento se desarrolló en dos etapas: en primer lugar, se determinó la cantidad óptima de componentes a retener, considerando aquellos que explican la mayor proporción de la varianza total del conjunto de datos. En segundo lugar, se analizó la composición de cada componente principal, con el fin de interpretar las variables que más contribuyen a su formación y entender la estructura subyacente de los datos.

Tal como se muestra en la Figura 12, el gráfico de sedimentación (Scree Plot) permite visualizar la proporción de varianza explicada por cada componente principal. A partir del análisis, se concluye que los dos primeros componentes son los más representativos, ya que explican en conjunto el 94,8 % de la varianza total. Además, se identifica un punto de inflexión claro en el segundo componente, lo que indica que la inclusión de un tercer o cuarto componente sería poco significativo para el análisis.

Figura 12.

Gráfico de Sedimentación (Scree Plot) - Calidad de Leches. Adaptado de resultados obtenidos en RStudio.



Fuente: Elaboración propia.

En la Tabla 6 se explica cómo esta compuesto cada uno de los componentes, explicaremos los elegidos previamente:

- PC1 (55.8% de varianza explicada): Este compuesto por valores positivos para cada una de las variables, con un mayor peso para las cuatro primeras (Grasa, Proteína, SNG, ST) símbolo de lotes de leche con calidad alta en todos sus aspectos en la composición. El pH no tiene influencia en el componente.
- PC2 (39.0% de varianza explicada): Este componente permitirá separar leches con alto contenido de grasa de aquellas que cuenten con un alto contenido de proteína. Fundamental bajo el entendimiento del negocio. Nuevamenta el pH no tiene influencia en el componente.

Tabla 6.

Análisis de carga de componentes - Variables de Calidad de Leche

Variable	**PC1**	**PC2**	**PC3**	**PC4**
Eigen	**2.17**	**1.603**	**0.227**	**0**
Grasa	-0.412	-0.627	0.065	0.658
Proteína	-0.522	0.421	0.742	0
SNG	-0.431	0.564	-0.623	0.33
ST	-0.61	-0.324	-0.239	-0.677

Fuente: Elaboración propia.

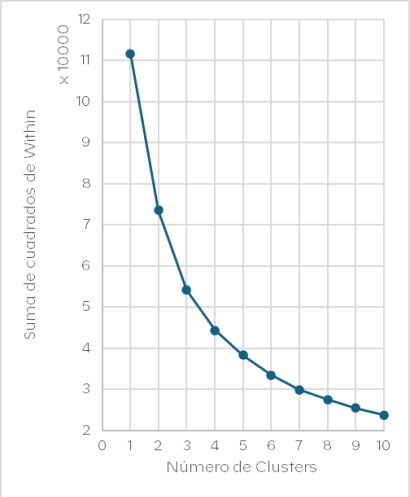
6.3. Determinación de número óptimo de clusters

La curva de suma de cuadrados intra-cluster que se muestra en la Figura 13 muestra una disminución pronunciada entre uno y tres clusters, y una reducción progresivamente menor a partir del cuarto. Este patrón indica la presencia de un 'codo' en k = 3 o 4, lo que sugiere que este rango proporciona una estructura de segmentación adecuada sin incurrir en una complejidad innecesaria. Por tanto, se seleccionó k = 3 para el modelo final.

El análisis del índice de silhouette que se muestra en la Figura 14 ratifica que el valor óptimo de clusters se encuentra en k = 3, dado que presenta el mayor nivel de cohesión interna y separación entre grupos.

Figura 13.

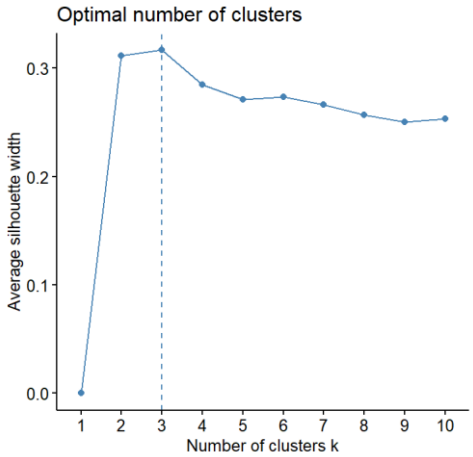
Cálculo de suma de Cuadrados intra-clusters



Fuente: Elaboración propia.

Figura 14.

Análisis del índice de Silhouette



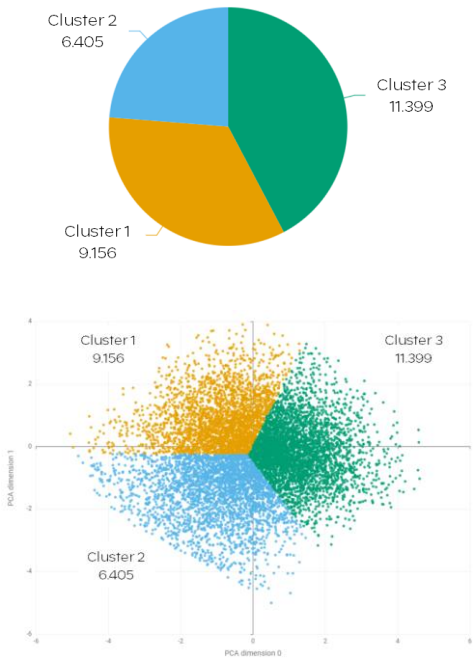
Fuente: Elaboración propia.

6.4. Resultados de clusterización

Una vez se ejecuta el modelo K-means, se obtiene la división de clústeres que se observa en la Figura 15. El gráfico evidencia una segmentación en tres grupos, generados mediante el algoritmo K-means. El Componente 1, interpretado como un indicador de calidad global (especialmente asociada a niveles de grasa y sólidos totales), permite distinguir un grupo ubicado hacia la derecha del eje, que corresponde a lotes con una menor calidad composicional. Por su parte, en el eje de las ordenadas, la componente 2 diferencia mejor la relación entre proteína y grasa, a valores superiores la proteína resulta mayor que la grasa y viceversa.

Figura 15.

Segmentación de Clusters - Participación y distriución en componentes.



Fuente: Elaboración propia.

Observando el diagrama de pie se observa que la mayor cantidad de datos se concentra en el Cluster 3 con 11.399 registros, el 42, lo cual indica que una gran parte del

acopio se encuentra en una zona de calidad estándar de leche. Sin embargo, los clústeres ubicados en la parte derecha del plano PCA capturan subgrupos diferenciados, que serán útiles para las decisiones operativas específicas, que serán expuestas a continuación.

6.5. Análisis de Clústeres y resultados de negocio

El modelo de segmentación mediante K-Means alcanzó convergencia en 5 iteraciones, lo cual indica una rápida estabilización de los centroides. El resultado final arrojó tres clústeres de tamaño desigual, y centroides con calidades diferentes que se resumen en la Tabla 7.

Media de las Propiedades de Calidad diferenciados por Clúster, se explican las características de cada clúster a continuación.

Tabla 7.

Media de las Propiedades de Calidad diferenciados por Clúster

Clúster	% de datos	n	Grasa	Proteína	SNG	ST
Clúster 1	34%	9,156	3.72	3.24	8.57	12.27
Clúster 2	24%	6,405	4.09	3.2	8.39	12.48
Clúster 3	42%	11,399	3.64	3.11	8.37	12.0

Fuente: Elaboración propia.

6.5.1. Clúster 1 – Leche Premium

Este primer clúster, agrupa un total de 9.156 registros, lo que representa aproximadamente el 34% del conjunto de datos analizado. Este clúster se distingue por presentar los mayores niveles de proteína y sólidos no grasos en comparación con los demás grupos identificados. Estas características de calidad posicionan a esta leche como la más adecuada para procesos de fabricación de queso, al concentrar una mayor proporción de sólidos útiles para la transformación industrial.

La relación entre el contenido de proteína de la leche y el rendimiento quesero resulta ser una correlación positiva. La proteína, en particular la caseína, constituye la fracción que se coagula y se retiene en la cuajada durante la elaboración del queso. En consecuencia, a medida que aumenta la concentración de proteína en la leche, se incrementa la cantidad de sólidos recuperados en el producto final, lo que se traduce en un mayor rendimiento de queso por unidad de leche procesada (kg de queso/100kg de leche). Diversos estudios señalan que leches con mayor contenido de proteína y sólidos no grasos permiten mejorar la eficiencia del proceso, reducir pérdidas de sólidos en el suero y generar impactos económicos positivos para la industria quesera. Por tanto, la clasificación de la leche Premium en un clúster diferenciado no solo responde a criterios estadísticos, sino que tiene una justificación tecnológica y económica directa, al asociarse con un mayor potencial de rendimiento industrial. (Walstra, Wouters, & Geurts, 2006)

Desde la perspectiva estratégica de una compañía láctea, este clúster representa una oportunidad significativa de reducción de costos, ya que cada incremento de 0,1% en el contenido de proteína de la leche tiene el potencial de disminuir el costo de producción del queso en más del 2%, como resultado de un mayor rendimiento industrial y una mejor eficiencia en el uso de la materia prima.

6.5.2. Clúster 2 – Leche como habilitador

El clúster “Leche como habilitador” agrupa 6.405 registros, equivalentes al 24 % del total analizado, y corresponde a la leche destinada a la elaboración de productos líquidos de la compañía, tales como leche entera, semidescremada, descremada y deslactosada y fabricaciones. Este clúster se caracteriza por presentar un mayor contenido de grasa en comparación con otros grupos, condición que resulta clave desde el punto de vista

operativo, ya que dicha fracción debe ser ajustada o estandarizada durante el proceso productivo para cumplir con las especificaciones de cada producto final.

Desde una perspectiva tecnológica e industrial, un mayor contenido de grasa en la leche permite la separación de crema mediante procesos de descremado. La crema obtenida constituye un insumo estratégico de alto valor, habilitando la fabricación de productos como mantequillas y cremas pasteleras, los cuales presentan una mayor rentabilidad en comparación con los productos líquidos básicos. La literatura especializada indica que la grasa láctea es un componente fundamental para el desarrollo de textura, sabor y funcionalidad en estos derivados, y su recuperación eficiente contribuye a una mejor valorización integral de la materia prima, maximizando el ingreso generado por litro de leche procesada (Walstra, Wouters, & Geurts, 2006)

En términos cuantitativos, el clúster "Leche como habilitador" adquiere una relevancia estratégica adicional al evidenciar que por cada incremento de 0,1 % en el contenido de grasa de la leche cruda, es posible obtener un aumento aproximado del 2,6 % en la cantidad de crema recuperable durante el proceso de descremado. Este incremento en la disponibilidad de crema permite alimentar de manera directa los procesos de mayor margen bruto, como la fabricación de mantequilla y cremas, maximizando la captura de valor a partir de la misma materia prima. En consecuencia, la gestión y segregación de leche con mayor contenido graso no solo optimiza la estandarización de productos líquidos, sino que se convierte en una ventana de oportunidad importante, al fortalecer la rentabilidad global del portafolio mediante una mayor producción de derivados con alto valor agregado.

6.5.3. Clúster 3 – Leche Estándar

Por último, se encuentra el Clúster 3, denominado "Leche estándar", que representa el 42 % de los registros, con un total de 11.399 datos. Se clasifica como leche estándar por

presentar una calidad adecuada para su uso en múltiples procesos productivos, funcionando como un insumo “colchón” que complementa o estabiliza la producción de otras líneas. Su valor agregado no es particularmente diferencial para la compañía, pero su versatilidad permite emplearla en fabricaciones generales, procesos de pulverización o como soporte en la producción de leche deslactosada o semidescremada, dado que cumple con la flexibilidad exigida por la normativa colombiana.

La segmentación obtenida permite a la organización comprender con mayor precisión la heterogeneidad del acopio de leche y aprovecharla estratégicamente en la distribución para fabricación de productos específicos que traen eficiencia operativa al proceso.

7. Modelado – Serie de Tiempo

Desde la perspectiva estratégica del proceso de acopio, resulta fundamental contar con la capacidad de proyectar el volumen de leche que ingresará a la planta en los días siguientes. Esta previsión permite optimizar la planificación operativa, anticipar necesidades de capacidad, ajustar los planes de producción y gestionar de forma más eficiente los recursos logísticos y de transformación. Con este propósito, se estableció un nuevo objetivo orientado a la predicción del acopio diario, para lo cual se aplicaron técnicas de series de tiempo y comparándolas para identificar el mejor ajuste en parámetros para su aplicación.

La selección de estos modelos se fundamenta en su idoneidad para capturar la dinámica temporal propia del acopio de leche, un proceso caracterizado por patrones estacionales, variaciones recurrentes asociadas a ciclos climáticos y comportamientos dependientes del tiempo. La elección de los modelos ARIMA y SARIMAX se soporta en literatura especializada.

La selección de un modelo ARIMA (AutoRegressive Integrated Moving Average) yace en su capacidad comprobada para capturar patrones de autocorrelación, tendencia y estacionalidad en series temporales univariadas sin requerir grandes cantidades de datos exógenos. Los modelos ARIMA integran componentes autorregresivos (AR), de media móvil (MA) y de diferenciación (I), lo que les permite adaptarse a estructuras dinámicas complejas y producir pronósticos robustos y ajustados a la variabilidad real de los datos históricos. Según Hyndman y Athanasopoulos, los modelos ARIMA constituyen una de las metodologías más utilizadas y eficaces en forecasting debido a que estimulan las relaciones temporales intrínsecas de la serie, minimizan errores de predicción y permiten construir

intervalos de confianza para cada pronóstico (Hyndman & Athanasopoulos, 2021). En el contexto de la planificación operativa de una compañía láctea, esta característica es esencial para generar alertas tempranas confiables, facilitar la toma de decisiones basadas en evidencia y sustituir métodos heurísticos previamente empleados.

Por su parte, la utilización de modelos extendidos del tipo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) resulta especialmente pertinente en contextos donde el comportamiento de la serie temporal está influenciado por factores externos observables. Estos modelos permiten incorporar variables exógenas que explican parte de la variabilidad no capturada únicamente por la estructura interna de la serie, como condiciones de mercado, factores climáticos, cambios operativos o restricciones logísticas. Durbin y Koopman (2012) señalan que la inclusión de variables exógenas dentro de un marco de modelos de espacio de estados mejora de manera significativa la capacidad predictiva y la interpretabilidad del modelo, al separar la dinámica inherente de la serie de los efectos causales externos. (Durbin & Koopman, 2012) En el contexto de la predicción de volúmenes de acopio en una compañía láctea, el enfoque SARIMAX permite capturar tanto la estacionalidad y autocorrelación del proceso como el impacto de variables exógenas relevantes, fortaleciendo la generación de pronósticos más precisos y operativamente accionables.

7.1. Preparación de datos

Para el desarrollo del modelado, se realizó el emparejamiento entre la base de datos de calidad de leche, compuesta por 26.960 registros, y la base de datos de volumen de acopio, que contiene 41.478 registros. En este proceso se tomó como estructura principal la base de volumen, de manera que, cuando no existía correspondencia con la base de calidad,

los valores faltantes se conservaron en blanco. Esta decisión responde a la necesidad de disponer posteriormente de la variable grasa como insumo exógeno en el modelo SARIMAX.

Una vez integradas ambas bases, se identificaron 14.764 registros con información faltante en alguna de las variables provenientes de la base de calidad. Para abordar esta situación, se implementó un proceso de imputación basado en la búsqueda del registro más cercano en términos de lote y fecha que contara con datos completos, utilizando dichos valores como reemplazo. Gracias a este procedimiento, los datos faltantes se redujeron a 619 registros. Finalmente, estos registros —equivalentes al 1,5 % del total— fueron eliminados, dado que no contaban con información suficiente para su uso en el análisis posterior.

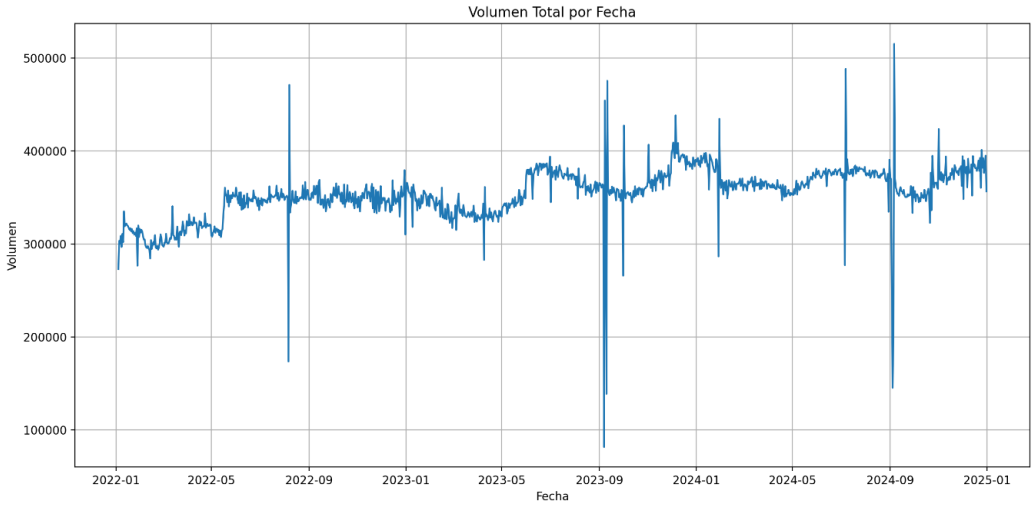
7.2. Análisis de datos históricos

Se inició por un análisis de la gráfica diaria de acopio de leche con los volúmenes que se tenían en la base de datos, recopilando información desde el 4 de enero de 2022 y hasta el 31 de diciembre de 2024. Obtenemos entonces una gráfica como la que se muestra en la Figura 16.

Se observa un comportamiento temporal continuo, donde los valores dependen de su posición en la secuencia, lo cual confirma la naturaleza cronológica del fenómeno y la necesidad de métodos específicos que capturen dependencias en el tiempo. Asimismo, la serie presenta una tendencia suave y ligeramente ascendente a lo largo del periodo analizado, iniciando sobre los 300.000L a inicios del año 2022 y llegando a los 380.000L para el final del 2024, reflejando un crecimiento moderado en la capacidad de acopio.

Figura 16.

Volumen de leche acopiada diariamente de 2022 a 2025



Fuente: Elaboración propia.

Adicional a lo anterior, es posible identificar una estacionalidad anual, expresada en fluctuaciones recurrentes que coinciden con patrones climáticos típicos del sector lácteo colombiano. Durante los periodos secos se evidencian disminuciones temporales del acopio, mientras que en temporadas lluviosas el volumen tiende a incrementarse debido a mejores condiciones de disponibilidad de alimento animal. Esta estacionalidad es clave para la planificación operativa y constituye un componente fundamental para el modelado con técnicas como SARIMA o SARIMAX, que permiten capturar variaciones cíclicas.

Finalmente, se aprecian caídas abruptas que corresponden a eventos puntuales de inestabilidad política en el país, conocidos como bloqueos en carreteras vividos hacia septiembre de 2023 y que se reflejan en una interrupción en el suministro de materia prima a la compañía, pero que, a su vez, en días siguientes, se observa un rebote positivo ingresando lo que en días anteriores se retuvo.

La combinación de tendencia, estacionalidad marcada y ruido justifica el uso de modelos de series de tiempo, ya que estos permiten capturar la estructura temporal del acopio, anticipar variaciones y soportar decisiones estratégicas relacionadas con la planeación de producción, logística y abastecimiento.

7.3. Prueba de estacionariedad

Se realizó la prueba de Dickey-Fuller Aumentada (ADF por sus siglas en inglés) en la que se obtuvo los siguientes resultados:

- ADF = -21.652
- p = 0
- Valor crítico:
    - 1% = -3.43
    - 5% = -2.86
    - 10% = -2.57

Dado que el valor *p* del test ADF es igual a 0 y el estadístico ADF es menor que todos sus valores críticos, se rechaza con alto nivel de confianza la hipótesis nula de presencia de raíz unitaria. Esto indica que la serie puede considerarse estacionaria, es decir, presenta una media, varianza y estructura de autocorrelación que se mantienen constantes a lo largo del tiempo. Por lo tanto, la serie cumple con una condición fundamental para la aplicación adecuada de modelos de series de tiempo.

7.4. MODELO ARIMA

El modelo ARIMA es una serie de tiempo que se compone por tres parámetros: p, d y q; los cuales describen la forma en como el modelo captura los datos. El parámetro p

representa el número de rezagos (valores pasados) de la serie que se utilizan para explicar su comportamiento actual; d indica el grado de diferenciación necesario para volver la serie estacionaria, es decir, cuántas veces deben restarse valores consecutivos para eliminar tendencias o variaciones no constantes; y q corresponde al número de términos de error rezagados que el modelo incorpora para corregir patrones no explicados directamente por los valores pasados

7.4.1. Selección de parámetros

Para el desarrollo del modelo ARIMA se llevaron a cabo las diferentes combinaciones para cada uno de los parámetros p, d y q. De 0 a 7 en cada uno de éstos y combinándolos, con lo anterior se generaron 343 iteraciones.

Adicional a lo anterior se tomaron BIC, AIC y el RMSE (Error cuadrado) como indicadores de éxito para la selección de parámetros de la serie ARIMA:

- AIC (Akaike Information Criterion) el cual mide la calidad del ajuste penalizando la complejidad del mismo. Se busca un valor bajo.
- BIC (Bayesian Information Criterion) genera un complemento al AIC en la búsqueda de modelos cuya eficiencia sea predominante sobre la complejidad. Se buscan valores bajos.
- RMSE (Root Mean Squared Error) evalúa el error del pronóstico de datos en validación. Se buscan valores bajos.

Tabla 8.

Comparativa de indicadores (AIC, BIC y RMSE) de los mejores 5 modelos ARIMA

Modelo (p,d,q)	AIC	BIC	RMSE
(3,5,5)	1.394	1.438	6.44×10^{14}
(3,0,3)	22.381	22.421	2.46×10^{4}
(3,0,2)	22.382	22.416	2.46×10^{4}

Fuente: Elaboración propia.

En la Tabla 8 se muestra los tres mejores modelos para el método ARIMA. Las diferencias entre estos se evidencian principalmente en el parámetro d, correspondiente al grado de diferenciación aplicado a la serie. Tal como lo exponen Hyndman y Athanasopoulos, valores excesivamente altos de d (mayores de 2) suelen indicar un sobreprocesamiento de la serie que elimina parte importante de su estructura temporal, generando inestabilidad numérica y pérdidas sustanciales de información. (Hyndman & Athanasopoulos, 2021) Esto se ve reflejado en los valores extremadamente altos de RMSE reportados para los modelos con mayor diferenciación, lo que demuestra que, aunque puedan presentar AIC o BIC bajos, su capacidad predictiva es muy deficiente. Por ello, dejar por fuera los modelos con altos valores de *d* es ventajoso, ya que permite conservar la naturaleza original de la serie y evita problemas de sobreajuste y degradación de la precisión del pronóstico.

Considerando estos resultados, los modelos con diferenciación equivalente a 0, específicamente (3,0,3) y (3,0,2), emergen como las alternativas más estables y confiables. Ambos presentan los valores más bajos de RMSE y un desempeño muy similar en términos de AIC y BIC; sin embargo, el modelo (3,0,2) resulta preferible debido a su menor complejidad, al utilizar un número más reducido de parámetros sin sacrificar calidad

predictiva. Por su equilibrio entre simplicidad, estabilidad y buen ajuste, este modelo se selecciona como la opción óptima para el proceso de modelado.

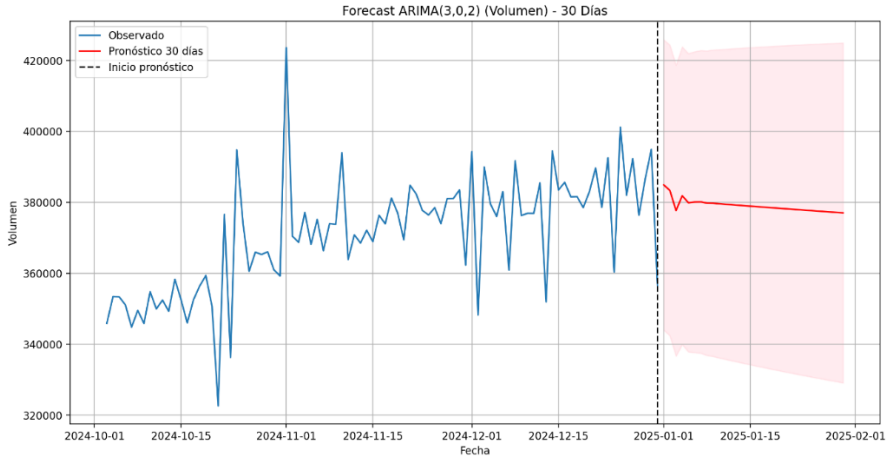
7.4.2. Pronóstico y análisis de negocio

En la Figura 17 se observa el pronóstico para el modelo ARIMA con parámetros (3,0,2). Se observa que el volumen de leche que ingresará al acopio durante los próximos 30 días se mantendrá relativamente estable, con ligeras variaciones alrededor de los valores recientes observados, más bien estables.

Esto es consistente con la naturaleza del modelo (d=0), que presupone que la serie no presenta tendencia significativa y que los patrones recientes se mantienen en el corto plazo. Adicionalmente, la banda de confianza (sombreado rosa) se expande progresivamente a medida que avanza el horizonte de predicción, lo cual es normal en modelos ARIMA y refleja el aumento natural de la incertidumbre conforme se proyectan fechas más alejadas.

Figura 17.

Pronóstico del modelo ARIMA (3,0,2)



Fuente: Elaboración propia.

. Desde la perspectiva de negocio, este comportamiento implica que el acopio puede esperar volúmenes similares a los que ha manejado en las semanas recientes, sin riesgos inmediatos de desabastecimiento ni expectativas de sobreoferta considerable. No obstante, las bandas de incertidumbre indican que se deben mantener márgenes de manejo adecuados, especialmente hacia el final del horizonte de 30 días, donde la variabilidad aumenta.

7.5. MODELO SARIMAX

Se escoge SARIMAX como un segundo modelado de serie de tiempo debido a que se considera una evolución del modelo SARIMA, ya que extiende sus capacidades al permitir la incorporación de variables exógenas que influyen sobre la serie temporal principal. Mientras SARIMA modela únicamente la estructura interna de la serie (tendencias, estacionalidades y autocorrelaciones propias), SARIMAX integra información adicional proveniente de factores externos que pueden explicar variaciones en el comportamiento del proceso.

En cuanto a su estructura, un modelo SARIMAX se caracteriza por la combinación de tres componentes: los parámetros ARIMA (p, d, q), que representan la autoregresión, el nivel de diferenciación y el modelo de medias móviles; los parámetros estacionales (P, D, Q, s), que modelan patrones repetitivos en periodos definidos.

Se tomará como variable exógena la Grasa al ser una variable que puede verse relacionada con los niveles de leche, y responder a estudio climatológicos que pueden verse reflejados en esta variable.

A continuación, se muestra el paso a paso para la selección del modelo SARIMAX.

7.5.1. Selección de parámetros

Basados en los resultados del modelo ARIMA, y por disminución de cálculo se lleva a cabo el proceso del resultado en la combinación de los parámetros de la siguiente manera:

- d y q se harán entre 0 y 3.
- D y Q se harán entre 0 y 2
- Los periodos se manejarán entre 7, 15 y 30 días.

Además, se toman los mismos indicadores que usamos en ARIMA para la selección de los parámetros óptimos BIC, AIC y RMSE (Ver explicación en sección 8.4.1.)

Tabla 9.

Comparativa de indicadores (AIC, BIC y RMSE) de los mejores 5 modelos SARIMAX

Modelo (p,d,q)	Estacional (P,D,Q,s)	AIC	BIC	RMSE
(2,1,2)	(1,1,1,30)	21.196	21.230	44.081
(2,1,2)	(0,1,1,30)	21.198	21.227	45.208
(1,1,2)	(1,1,1,30)	21.241	21.270	45.754
(1,0,2)	(1,1,1,30)	21.246	21.275	29.170
(1,1,2)	(0,1,1,30)	21.246	21.270	48.030

Fuente: Elaboración propia.

Al comparar los valores de RMSE, se observa que existe una diferencia notable entre los modelos. Aunque el modelo (2,1,2)(1,1,1,30) presenta el RMSE más bajo entre las primeras alternativas (44.081), el valor más destacado se obtiene en el modelo (1,0,2)(1,1,1,30) con RMSE equivalente a 29.170.

Esta diferencia es sustancial y evidencia que el modelo (1,0,2)(1,1,1,30) es significativamente más preciso en términos predictivos. Dicho modelo logra reducir el error de pronóstico de forma considerable en comparación con los demás, mostrando una mejor capacidad para capturar la dinámica real del proceso aun con un componente no diferenciado en la parte regular (d = 0). Esto sugiere que la estacionalidad y los términos

AR y MA ya explican adecuadamente la estructura temporal sin necesidad de una diferenciación adicional, lo que además preserva estabilidad y evita una pérdida innecesaria de información.

Por su parte, la componente estacional, (1,1,1,30) muestra que existe una dependencia clara entre el comportamiento del volumen actual y el de un ciclo anterior de aproximadamente 30 días, lo cual es coherente con la variación mensual típica en la producción láctea.

7.5.2. Pronóstico y análisis de negocio

En la Figura 18 se observa el pronóstico para el modelo SARIMAX con parámetros (1,0,2)x(1,1,1,30) y grasa como variable exógena. Se observa que el volumen de leche que ingresará al acopio durante los próximos 30 días se mantendrá relativamente estable, con ligeras variaciones alrededor de los valores cercanos a los 380.000L diarios.

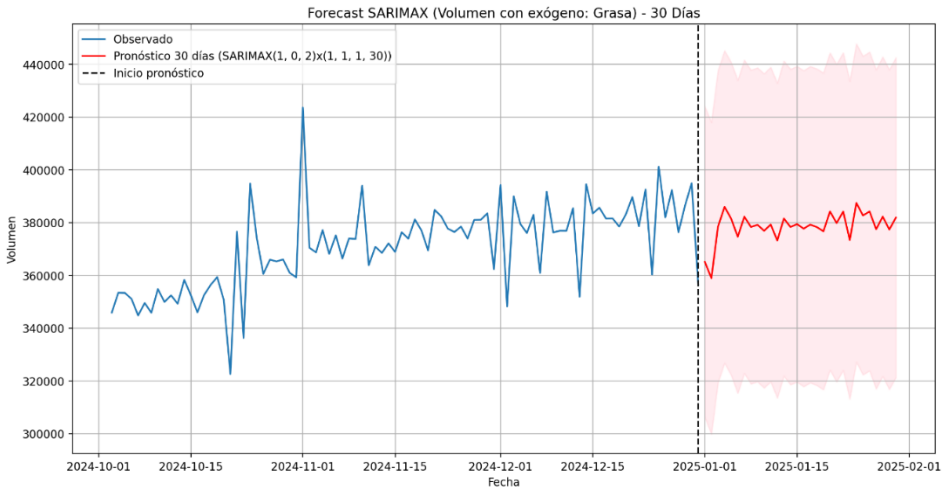
Esto es consistente con la naturaleza del modelo (d=0), que presupone que la serie no presenta tendencia significativa y que los patrones recientes se mantienen en el corto plazo. Adicionalmente, la banda de confianza (sombreado rosa) se expande progresivamente a medida que avanza el horizonte de predicción, lo cual es normal en modelos ARIMA y refleja el aumento natural de la incertidumbre conforme se proyectan fechas más alejadas.

Desde una perspectiva de negocio, este pronóstico valida la capacidad del modelo para capturar la estacionalidad mensual (s=30) y la tendencia reciente, ofreciendo a la gerencia de operaciones un escenario base de estabilidad relativa para la planificación logística y de almacenamiento. No obstante, la amplitud de los intervalos de confianza (zona sombreada roja) sugiere una volatilidad latente que no debe ignorarse; por tanto,

aunque se espera un flujo constante, se recomienda mantener una capacidad de respuesta flexible en el centro de acopio para gestionar picos potenciales que podrían superar los 440.000 litros, o valles inferiores a los 320.000, asegurando así la continuidad operativa sin incurrir en sobrecostos por capacidad ociosa o desbordamiento.

Figura 18.

Pronóstico del modelo SARIMAX (1,0,2)x(1,1,1,30)



Fuente: Elaboración propia.

7.6. Comparación de modelos

Con la finalidad de llevar a cabo un ejercicio comparativo entre los modelos con los valores reales obtenidos para el mes de enero del presente año, se presenta la Figura 19 con la gráfica de volumen para la cual no se presentan las dimensiones del eje vertical por confidencialidad. La gráfica presenta una comparación del volumen real (línea continua azul oscuro) con las predicciones generadas por los modelos de series de tiempo, ARIMA (línea punteada azul claro) y SARIMAX (línea discontinua gris), a lo largo de 30 días siguientes. El volumen real muestra una alta volatilidad, caracterizada por picos extremos en los día 6 y 7 indicando la presencia de valores atípicos significativos. En contraste, el

modelo ARIMA ofrece una predicción notablemente suave y estable, actuando como un pronóstico de media a largo plazo que no logra capturar la varianza a corto plazo ni los picos.

Figura 19.

Gráfica de los valores de volumen real y pronósticos ARIMA y SARIMAX para los 30 días del mes de enero 2025

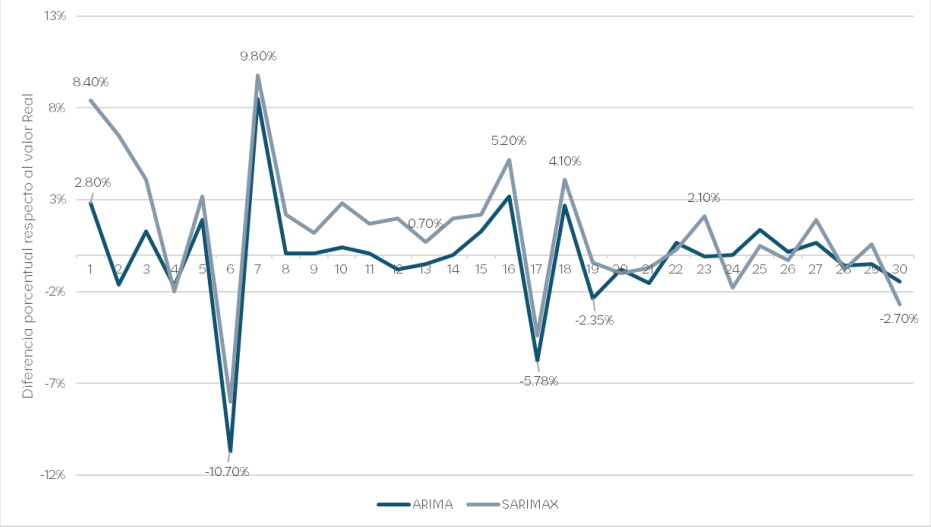
Fuente: Elaboración propia.

Por su parte, el modelo SARIMAX demuestra una mejor capacidad de ajuste a las fluctuaciones del volumen real, especialmente en la captura de la tendencia general y la volatilidad, aunque con un evidente desfase temporal. Si bien la línea SARIMAX logra replicar las direcciones de subida y bajada (como alrededor de los períodos 1-6 y 17-27), su capacidad predictiva se ve limitada por la intensidad de la volatilidad, lo que podría indicar que la grasa como variable exógena en la estructura de estacional fue efectiva para modelar la varianza.

Para un mayor entendimiento de lo anterior expuesto, en la Figura 20 se muestra el diagrama porcentual de la estimación. Siendo la línea base de 0% el valor real que se obtuvo en las 30 observaciones siguientes.

Figura 20.

Diferencia porcentual de los valores obtenidos por los modelos de ARIMA y SARIMAX en comparación con el descargue real.



Fuente: Elaboración propia.

Se observa como la magnitud de los cambios en los datos reales aún presenta un desafío. En términos de sesgo, ambos modelos tienden a subestimar los picos y sobreestimar los valles extremos con valores de hasta el 10,7% de subestimación para un día.

Desde el entendimiento del negocio un 10% de estimación puede referirse a 40.000L de leche que si bien pueden afectar un día, se observa que al siguiente se recuperan y no afectan en la logística de la empresa dichas variaciones diarias que se anulan entre días contiguos, toda vez que se cuenta con un stock de leche representativo dentro de la compañía que permite acolchonar estas diferencias puntuales.

Por último, si llevamos a cabo la sumatoria de los volúmenes obtenidos para la estimación de los 30 días se observa que los modelos cuentan con una diferencia porcentual

respecto al volumen acumulado acopiado de -0.027% para el modelo ARIMA y -0.127% para el modelo SARIMAX. Siendo en ambos casos, estimaciones que no exceden diferencias por más de 15.000L de leche en un acopio que asciende los 10 millones de litros mensuales, valores excelentes para un pronóstico.

8. Plan y recomendaciones de implementación y aplicación

Inicialmente, los modelos son utilizados por un equipo de líderes que dan las instrucciones de separación de leches a la operación según los datos obtenidos, lo que ha permitido capitalizar ahorros en el corto plazo en la compañía.

Es importante aclarar que en términos de tecnología, conocimiento e infraestructura se cuenta con las herramientas necesarias desde la informática a la operación para la capitalización de ahorros mediante la aplicación de las soluciones planteadas, llevando el proyecto a una fase de cero inversión para esta primera fase.

Para dar continuidad y puesta en marcha al presente proyecto empresarial se recomiendan los siguientes pasos:

- Capacitación del personal de trabajo que utilizará el tablero operativo para la separación de leches por calidad.
- Capacitación del personal de trabajo que utilizará el tablero estratégico para el Forecast de volúmenes de leche acopiados.
- Actualizar la información para el modelo de Forecast de manera mensual, para así poder identificar y anticipar cambios en los volúmenes que se puedan presentar en la región debido a cambios climáticos o en el proceso logístico de acopio.
- Actualizar de manera mensual el modelo de clusterización de tal manera que la información se mantenga ajustada a los cambios en clima y tener la posibilidad de mantener una precisión en el uso de leches dentro de la compañía.

9. Conclusiones

1. La implementación de modelos analíticos de segmentación de la leche por calidad permitió optimizar su asignación dentro de los diferentes procesos productivos de la compañía, generando eficiencias operativas medibles y contribuyendo directamente al incremento del margen mediante un uso más estratégico de la materia prima. De forma complementaria, el modelo de series de tiempo para la predicción de volúmenes de acopio proporcionó una herramienta robusta para la anticipación de escenarios operativos, habilitando la generación de alertas tempranas que fortalecen la planificación productiva y reemplazan prácticas previamente basadas en estimaciones empíricas o especulativas.
2. La implementación de la metodología CRISP-DM sirvió como un marco robusto y estructurado, permitiendo una transición efectiva desde el entendimiento del negocio, hasta el despliegue de los modelos. Este enfoque garantizó que los esfuerzos de análisis estuvieran directamente alineados con los objetivos de rentabilidad y eficiencia operativa de la compañía.
3. La combinación de Análisis de Componentes Principales (PCA) y el algoritmo K-means permitió reducir la dimensionalidad y segmentar eficazmente la leche acopiada en tres clústeres óptimos (Leche Premium, Leche para Leche y Leche Estándar) basada en el perfil composicional (grasa, proteína, SNG, ST) proporcionando una base para la asignación del eficiente del insumo basado en datos.
4. Aunque el modelo ARIMA (3,0,2) evidenció un comportamiento estable, el modelo SARIMAX (1,0,2)×(1,1,1,30), que incorpora la grasa como variable exógena,

demostró un desempeño superior al obtener el RMSE más bajo entre las alternativas evaluadas. Aun cuando se observó una volatilidad diaria considerable (con errores cercanos al 10.7%), ambos modelos presentaron un margen de error mínimo en el pronóstico del volumen total acumulado a 30 días, con desviaciones inferiores al 0.13%.

10. Bibliografía

Christopher, M., & Holweg, M. (2011). Supply Chain 2.0: Managing supply chains in the era of turbulence. *International Journal of Physical Distribution & Logistics Management*, 63-82.

Durbin, J., & Koopman, S. (2012). *Time Series Analysis by State Space Methods (2nd ed.).* Oxford University Press.

FAO. (2025). *Gateway to dairy production and products: Milk composition*. Obtenido de https://www.fao.org/dairy-production-products/products/milk-composition/en

Fedegan. (2024). *Federación Colombiana de Ganaderos*. Obtenido de Consumo aparente per cápita anual Leche: https://www.fedegan.org.co/estadisticas/consumo-0

Fox,, P., & McSweeney, P. (2015). *Dairy chemistry and biochemistry.* Springer. doi:10.1007/978-1-4419-8602-3

Haug, A., Høstmark , A. T., & Harstad, O. M. (2007). Bovine milk in human nutrition – a review. *Lipids in Health and Disease*, 1-8.

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd ed.).* OTexts.

Kourkouta et al. (2021). Milk Nutritional Composition and Its Role in Human. *Journal of Pharmacy and Pharmacology 9*, 10-15.

OCLA. (2023). *Consumo Mundial per cápita y población.* Obtenido de https://www.ocla.org.ar/noticias/10015011-consumo-mundial-per-capita-y-poblacion

Our World in Data. (2022). *Per capita milk consumption [Mapa interactivo]*. Obtenido de OurWorldInData.org: https://ourworldindata.org/grapher/per-capita-milk-consumption?time=2022

Our World in Data. (2023). *Our World in Data*. Obtenido de Milk Production: https://ourworldindata.org/grapher/milk-production-tonnes

Raddar; Asoleche; Ministerio de Agricultura. (2025). *Lácteos en la mesa de los colombianos.* Bogotá, Colombia.

Walstra, P., Wouters, J., & Geurts, T. (2006). *Dairy science and technology.* Boca Raton: CRC Press.