UNIVERSIDAD
NACIONAL
DE COLOMBIA

Optimización del pronóstico de ventas en Dislicores: Evaluación y selección de modelos de series temporales, aprendizaje automático y técnicas avanzadas.

Wilson Fernando Espitia Humanez

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2025

Optimización del pronóstico de ventas en Dislicores: Evaluación y selección de modelos de series temporales, aprendizaje automático y técnicas avanzadas.

Wilson Fernando Espitia Humanez

Tesis Presentada como requisito parcial para optar al título de:

Magister en Ingeniería – Analítica

Director (a):

Ph.D., María Constanza Torres Madroñero

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2025

Dedicatoria

A mis padres, Rafael y Victoria, a mis hermanos: José Manuel, Ana Cecilia y Ana Karina

Agradecimientos

Quiero expresar mi más profundo agradecimiento a toda mi familia, especialmente, a mis padres, Rafael y Victoria. Su carácter, disciplina y amor han forjado en mí una voluntad indómita fundamentada en la ética, la moral y el respeto hacia las ciencias, el conocimiento, la razón y la justicia. Gracias a su excelente crianza, he podido desarrollarme como un hombre razonable, pensante y culto.

Mi gratitud también se extiende a la empresa Dislicores, por proporcionar un ambiente propicio, las herramientas necesarias y el apoyo indispensable que me permitieron transformar esta propuesta en un proyecto tangible y exitoso.

De manera especial, agradezco a la Doctora María Constanza Torres Madroñero, por su excelente orientación, acompañamiento y guía durante todo el desarrollo de esta tesis. Su profesionalismo ha dejado una huella en mi formación académica.

Finalmente, agradezco profundamente a Dios por el don del conocimiento y la sabiduría, y a la Universidad Nacional de Colombia por brindarme la oportunidad de aplicar y compartir este conocimiento en beneficio de la sociedad.

Resumen

Optimización del pronóstico de ventas en Dislicores: Evaluación y selección de modelos de series temporales, aprendizaje automático y técnicas avanzadas.

Dislicores es una empresa colombiana líder en la distribución y comercialización de bebidas alcohólicas y no alcohólicas, siendo su portafolio principal los vinos (tintos, blancos, rosados y espumosos), licores (whisky, ginebra, rones y tequilas), cervezas y bebidas sin alcohol. Además, complementa su oferta con alimentos como pasabocas, dulces y alimentos madurados, fortaleciendo y complementando su propuesta de consumo masivo.

El objetivo de este estudio es mejorar la precisión de los pronósticos de demanda, mediante la evaluación y comparación de modelos avanzados de predicción. Para ello, se implementan seis modelos: Promedio Móvil Simple, Regresión Lineal, Holt-Winters, MSTL con Regresión Polinómica, XGBoost y LightGBM, utilizando datos históricos del periodo 2020-2024.

Siguiente la metodología CRISP-DM, se realizaron procesos de limpieza de datos, análisis exploratorio y evaluación de desempeño empleando métricas como MAE%, SESGO%, SCORE% y RMSE% sobre las 232 referencias clave (Pareto). El modelo seleccionado se integrará en un pipeline automatizado y escalable dentro del servicio en nube de Amazon Web Services (AWS).

Finalmente, los resultados se implementaron en Qlik Sense, facilitando el análisis y toma de decisiones estratégicas basadas en datos confiables y de fácil acceso. Esta solución pretende contribuir a una gestión y seguimiento más eficiente de la demanda, optimizando la planificación operativa y fortalecer la competitividad de Dislicores en el mercado nacional.

Palabras clave: Pronóstico de demanda, series temporales, aprendizaje automático, modelos Holt-Winters, XGBoost, LightGBM, gestión de inventarios, análisis predictivo, Business Intelligence.

Abstract

Sales forecast optimization in Dislicores: evaluation and selection of time series models, automated learning and advanced techniques.

Dislicores is a leading Colombian company in distribution and marketing of alcoholic and non-alcoholic beverages, its main portfolio being wines (red, white, rosé and sparkling), liquor (whiskey, gin, rum, tequila), beers and non-alcoholic drinks. Also, complement its offer with foods as snacks, candies and aged foods strengthening and complementing their massive consumption offer.

This study's focus is to improve demand forecast accuracy through evaluation and comparison of advanced forecasting models. To do this, six models are implemented: Simple Moving Average (SMA), Linear Regression, Holt-Winters, MSTL with Polynomial Regression, XGBoost and LightGBM, using historical data from 2020 to 2024.

Then CRISP-DM methodology. Data cleaning, exploratory data analysis and performance evaluation processes were carried out using metrics like MAE, SESGO, SCORE, RMSE on the 232 key references (Pareto). The selected model will be integrated in an automated and scalable pipeline on the Amazon Web Services (AWS) cloud.

Finally, results were implemented in Qlik Sense, easing analysis and strategic decision-making based on reliable and easily accessible data. This solution aims to contribute to more efficient demand management and monitoring, optimizing operational planning and strengthening Dislicores' competitiveness in the national market.

Keywords: Demand forecasting, time series, automated learning, Holt-Winters models, XGBoost, LightGBM, inventory management, predictive analytics, Business Intelligence.

Contenido

Resumen 5
Abstract 6
Lista de figuras 10
Lista de tablas 10
1. Introducción 11
2. Objetivos 14
2.1. Objetivo general 14
2.2. Objetivos específicos 14
3. Marco teórico 15
3.1. Fundamentos del pronóstico de demanda 16
3.2. Importancia del pronóstico en la gestión empresarial 16
3.3. Desafíos en el pronóstico de la demanda en el sector de bebidas 17
3.4. Series temporales y modelos tradicionales 18
3.4.1. Conceptos básicos de series temporales 18
3.4.2. Modelo Holt-Winters 19
3.4.3. Regresión lineal simple 21
3.4.4. Promedio móvil simple 22
3.4.5. MSTL con regresión polinómica 22
3.4.6. XGBoost 23
3.4.7. LightGBM 25
3.4.8. Comparativa de modelos 27
3.5. Métricas de evaluación para los pronósticos de los modelos. 30
3.5.1. MAE% (Error Absoluto Medio Porcentual) 31
3.5.2. SESGO% (Porcentaje de Sesgo) 31
3.5.3. SCORE% 32
3.5.4. RMSE% (Raíz del error cuadrático medio) 33
3.6. Revisión de la literatura 33
4. Metodología implementada 35
4.1. Diseño de la investigación 36
4.1.1. Enfoque cualitativo y cuantitativo 36
4.1.2. Justificación del uso de la metodología CRISP-DM 38

4.2. Fases de la metodología CRISP-DM 39
4.2.1. Fase 1: Comprensión del negocio 40
4.2.2. Fase 2: Comprensión de los datos 41
4.2.3. Fase 3: Preparación de los datos 42
4.2.4. Fase 4: Modelado 42
4.2.5. Fase 5: Evaluación 44
4.2.6. Fase 6: Despliegue 45
4.3. Consideraciones éticas y legales 46
5. Análisis exploratorio de datos (EDA) 47
5.1. Descripción de los datos 47
5.1.1. Características generales de los datos de ventas 48
5.1.2. Variables estacionales y exógenas consideradas 51
5.2. Validación de los datos 52
5.2.1. Validación de limpieza de datos para la dimensión SKU 52
5.2.2. Validación de limpieza de datos para la dimensión Ítem 53
5.2.3. Validación de limpieza de datos para la tabla de hechos FactVenta 54
6. Implementación y comparación de modelos 61
6.1. Modelos de series de tiempo 61
6.1.1. Modelo de línea base actual: Proceso Supply Chain como Benchmark 62
6.1.2. Supply Chain 63
6.1.3. Regresión línea Supply Chain 64
6.1.4. Holt Winters Supply Chain 65
6.1.5. Promedio Móvil Simple 66
6.1.6. MSTL con Regresión Polinómica 67
6.1.7. Holt-Winters 67
6.2. Modelos de machine learning 68
6.2.1. Definición de búsqueda de Hiperparámetro 69
6.2.2. Generación de atributos 69
6.2.3. XGBoost local 70
6.2.4. XGBoost global 72
6.2.5. LigthGBM local 73
6.2.6. LigthGBM global 74
6.3.1. Evaluación global del forecast 76

6.3.2. Mejor modelo 76
6.3.3. Distribución de mejor modelo por SKU 78
6.3.4. Modelo con peor desempeño 79
6.3.5. Versus: Modelos SC (Supply Chain) vs Modelos optimizados 79
7. Implementación y selección de modelos 82
7.1. Resultado de ejecución periodo 2025-01 86
7.1.1. SKU 106009 87
8. Desarrollo de arquitectura tecnológica 94
9. Conclusiones 97

Lista de figuras

Figura 1. Esquema de XGBoost. 25
Figura 2.Algoritmo Boosting de machine learning. 26
Figura 3.Metodología CRISP-DM y sus fases. 38
Figura 4.Diagrama de arquitectura del despliegue del modelo de pronóstico. 45
Figura 5.Ventas totales por año. 55
Figura 6. Ventas promedio por mes. 55
Figura 7.Top 3 productos con más ventas por año. 56
Figura 8.Comportamiento de ventas por mes y año. 57
Figura 9.Descomposición, tendencia, estacionalidad y ruido. 58
Figura 10.Serie temporal de ventas mensuales. 59
Figura 11.Tendencia de las ventas (STL). 59
Figura 12.Ruido o residuo de las ventas (STL). 60
Figura 13.Distribución de mejor modelo por SKU. 78
Figura 14.Serie de tiempo: Demanda y pronóstico por SKU 84
Figura 15.Serie de tiempo: Demanda y pronóstico seleccionado por SKU. 84
Figura 16.Serie temporal de top3 SKU. 87
Figura 17. Serie temporal de demanda histórica real – SKU 106009 resaltando demanda del 2025. 88
Figura 18.. Serie temporal de proyección de demanda – SKU 106009. 91
Figura 19.Serie temporal de demanda histórica más proyección de ventas – SKU 106009. 92
Figura 20.Arquitectura de flujo en AWS. 94

Lista de tablas

Tabla 1.Comparativa de modelos de pronósticos: 28
Tabla 2.Especificación de la tabla FactVenta. 49
Tabla 3.Comparación de modelos optimizados vs Modelo SC. 80
Tabla 4. Tabla de Top3 mayor ventas de SKU’s entre 2020-01 y 2024-12. 86

1. Introducción

El entorno industrial se caracteriza por su alta competitividad y dinamismo. En este contexto, la precisión en las proyecciones de demanda se ha convertido en un factor crítico para el éxito, la sostenibilidad y la competitividad de las organizaciones (Chopra & Meindl, 2016). Las empresas deben enfrentarse a mercados cada vez más volátiles y a consumidores más exigentes, lo que exige estrategias adaptativas y basadas en datos precisos y confiables para respaldar la toma de decisiones informadas (Kotler & Keller, 2016).

Dislicores, una empresa colombiana líder en la distribución y comercialización de licores nacionales e importados, vinos, cervezas, bebidas no alcohólicas y alimentos complementarios como pasabocas, dulces y alimentos madurados, no es ajena a estos desafíos. La compañía ha identificado la necesidad de mejorar la exactitud de sus pronósticos de ventas como una medida esencial para mantener su posición competitiva en el mercado.

La importancia de contar con proyecciones de demanda precisas radica en la optimización de recursos, la reducción de costos y la mejora en la toma de decisiones estratégicas (Mentzer & Moon, 2005). Proyecciones inexactas pueden conducir a excesos o faltantes de inventario, afectando negativamente la rentabilidad y la satisfacción del cliente (Silver, Pyke, & Thomas, 2017). En el caso de Dislicores, las metodologías tradicionales empleadas hasta el momento, como proyecciones basadas en promedios móviles simples o juicio experto, no han capturado adecuadamente las complejidades y tendencias del mercado, especialmente considerando factores estacionales y eventos especiales que afectan el consumo de bebidas (Hyndman & Athanasopoulos, 2018).

Además, el sector de bebidas alcohólicas y no alcohólicas está sujeto a variaciones estacionales significativas, influenciadas por festividades, condiciones climáticas y campañas promocionales (Gupta & George, 2016). Estos factores aumentan la complejidad del pronóstico y requieren de modelos más sofisticados que puedan capturar patrones no lineales y relaciones complejas entre variables (Makridakis, Spiliotis, & Assimakopoulos, 2018). La incapacidad de anticipar adecuadamente estos cambios puede resultar en

oportunidades perdidas o en costos adicionales por almacenamiento excesivo o roturas de stock (Christopher, 2016).

Este estudio surge de la necesidad de resolver esta problemática, buscando soluciones que mejoren la precisión en las proyecciones de demanda de Dislicores. Para ello, se propone evaluar y seleccionar el modelo de pronóstico más adecuado, utilizando una combinación de modelos clásicos y avanzados como el Promedio Móvil Simple, Regresión Lineal, Holt-Winters, MSTL con regresión polinómica, y técnicas modernas de aprendizaje automático como XGBoost y LightGBM (Chen & Guestrin, 2016; Ke et al., 2017; Hyndman, Wang & Laptev, 2021). La combinación de estos modelos busca aprovechar las fortalezas de cada enfoque y mejorar la capacidad predictiva de la solución (Makridakis et al., 2018).

La integración de modelos de series temporales con algoritmos de aprendizaje automático ofrece un enfoque híbrido capaz de capturar tanto patrones históricos como relaciones no lineales y complejas en los datos (Zhang, 2003). Por ejemplo, los modelos basados en árboles de decisión como XGBoost y LightGBM han demostrado un excelente rendimiento en tareas de predicción debido a su capacidad para manejar grandes volúmenes de datos y capturar interacciones entre variables (Chen & Guestrin, 2016; Ke et al., 2017). Al aplicar estos modelos al pronóstico de ventas, es posible mejorar la precisión y adaptabilidad de las proyecciones (Makridakis et al., 2018).

El objetivo principal de este estudio es mejorar la precisión de las proyecciones y optimizar la gestión de la demanda y la planificación estratégica de Dislicores. Al profundizar en el análisis de los datos históricos de ventas del período 2020 a 2024, se busca comprender mejor las tendencias, patrones estacionales y posibles anomalías que puedan afectar el comportamiento de la demanda (Hyndman & Athanasopoulos, 2018). La implementación y comparación de diversos modelos de pronóstico permitirá identificar aquellos que ofrecen un mejor desempeño en términos de métricas como el MAE%, SESGO% y SCORE% (Armstrong, 2001).

Asimismo, se reconoce la importancia de desarrollar una arquitectura tecnológica sostenible y escalable que facilite la integración de los modelos seleccionados en la práctica diaria de la empresa. La utilización de servicios en la nube como Amazon Web Services (AWS) permite gestionar grandes volúmenes de datos y realizar cálculos complejos de manera eficiente (Amazon Web Services, 2021). La integración de los resultados en una plataforma

de Business Intelligence como Qlik Sense garantizará que el personal de Dislicores pueda acceder fácilmente a la información y utilizarla para la toma de decisiones (Qlik, 2021).

El alcance de esta investigación se centra en las ventas históricas de Dislicores durante el período 2020 a 2024, enfocándose en las 232 referencias de mayor impacto (referencias Pareto), las cuales representan un porcentaje significativo de las ventas totales (Pareto, 1971). Se considerarán variables estacionales y exógenas relevantes, como festividades, eventos promocionales y factores económicos que puedan influir en el comportamiento de la demanda (Fildes & Goodwin, 2007). La inclusión de estas variables es crucial para mejorar la precisión de los modelos y reflejar de manera más realista las condiciones del mercado (Song & Li, 2008).

Entre las limitaciones del estudio, se reconoce la dependencia de la calidad y disponibilidad de los datos. Datos incompletos, inconsistentes o con errores pueden afectar la precisión de los modelos y conducir a conclusiones erróneas (Kim, Trimi, & Chung, 2014). Además, el mercado puede experimentar cambios abruptos o eventos imprevistos, como crisis económicas o cambios en la legislación, que no pueden ser anticipados por los modelos estadísticos (Taleb, 2007). No obstante, se espera que los hallazgos de esta investigación aporten un valor significativo tanto para Dislicores como para el campo de la ingeniería analítica en general.

En términos metodológicos, se adopta la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), que proporciona un marco estructurado para llevar a cabo proyectos de minería de datos y análisis predictivo (Wirth & Hipp, 2000). Este enfoque garantiza una comprensión profunda del negocio y los datos, así como un proceso iterativo de modelado y evaluación que mejora la calidad de los resultados (Chapman et al., 2000).

La relevancia de este estudio radica en su contribución al avance del campo de la ingeniería analítica y la gestión de la cadena de suministro. Al combinar técnicas tradicionales con algoritmos modernos de aprendizaje automático, se ofrece una perspectiva innovadora que puede ser aplicada en otros contextos y sectores (Choi, Chan, & Yue, 2021). Además, el desarrollo de una arquitectura tecnológica sostenible y escalable en AWS facilita la implementación práctica y la adaptación futura a nuevas necesidades o cambios tecnológicos (Amazon Web Services, 2021).

En conclusión, esta investigación aborda los desafíos actuales que enfrenta Dislicores en la precisión de sus proyecciones de demanda, proponiendo una solución integral que combina modelos avanzados de pronóstico con una arquitectura tecnológica robusta. Los resultados esperados no solo beneficiarán a la empresa al mejorar su eficiencia operativa y estratégica, sino que también contribuirán al conocimiento y práctica en el campo de la ingeniería analítica y la gestión de datos. Al mejorar la precisión de los pronósticos, Dislicores podrá optimizar sus niveles de inventario, reducir costos y responder de manera más efectiva a las necesidades del mercado, fortaleciendo así su posición competitiva (Mentzer & Moon, 2005).

2. Objetivos

2.1. Objetivo general

Evaluar y seleccionar el modelo de pronóstico de ventas más adecuado para la empresa Dislicores, mediante la implementación y comparación de modelos de series temporales y aprendizaje automático, específicamente: Promedio Móvil Simple, Regresión Lineal, Holt-Winters, MSTL con regresión polinómica, XGBoost y LightGBM, considerando variables estacionales y exógenas relevantes durante el período 2020-2024, con el fin de mejorar la precisión de las proyecciones de ventas y optimizar la gestión del inventario y la planificación estratégica comercial.

2.2. Objetivos específicos

1. Realizar un análisis exploratorio de los datos históricos de ventas de Dislicores del período 2020-2024, comprendiendo su estructura, calidad y tendencias, e identificando y corrigiendo posibles anomalías o irregularidades, para garantizar la fiabilidad de los datos para el análisis y modelado posteriores.

2. Implementar y comparar múltiples modelos de pronóstico, incluyendo Promedio Móvil Simple, Regresión Lineal, Holt-Winters, MSTL con regresión polinómica, XGBoost y LightGBM, utilizando métricas de evaluación como MAE%, SESGO%,

SCORE% y RMSE%, con el objetivo de determinar cuál modelo presenta mejor desempeño para las 232 referencias clave (referencias Pareto) seleccionadas.

3. Seleccionar y optimizar el modelo más adecuado, ajustando sus parámetros para mejorar su capacidad predictiva y optimizar la exactitud del pronóstico de ventas de Dislicores, lo que garantizará una toma de decisiones más precisa y fundamentada en la gestión del inventario y la planificación estratégica comercial.

3. Marco teórico

Los pronósticos empresariales son herramientas fundamentales en la planificación estratégica de cada organización, ya que permiten asignar eficientemente los recursos, identificar nuevas oportunidades de negocio y mitigar riesgos en un entorno dinámico y competitivo (Chopra & Meindl, 2016). Según Mentzer y Moon (2005), un pronóstico preciso ayuda a las organizaciones a reducir la incertidumbre inherente al mercado, facilitando decisiones fundamentadas en datos y no en suposiciones.

Un ejemplo destacado es el caso de Nutresa, que utiliza pronósticos de demanda para planificar su producción y distribución, garantizando la entrega oportuna a sus clientes. Este enfoque no solo optimiza los recursos de producción, sino que también mejora la eficiencia en la logística, contribuyendo a su sostenibilidad operativa. De igual forma, empresas como Grupo Éxito han implementado modelos avanzados de predicción, logrando evitar tanto la escasez como el exceso de inventario, y asegurando así una gestión más eficiente de sus operaciones (Silver, Pyke, & Thomas, 2017).

3.1. Fundamentos del pronóstico de demanda

El pronóstico de demanda es una herramienta esencial para la gestión empresarial, ya que permite anticipar la cantidad de productos o servicios requeridos en el futuro, facilitando la toma de decisiones estratégicas en áreas como producción, inventarios y logística. Aunque los pronósticos no son infalibles, proporcionan una base sólida para reducir la incertidumbre y optimizar la planificación (Jain, 2012).

De acuerdo con Jain (2012), la clave de un pronóstico exitoso radica en la combinación de modelos estadísticos, análisis de datos y conocimiento del negocio. Esto permite a las organizaciones anticiparse a fluctuaciones del mercado y ajustar sus estrategias para mejorar la eficiencia y minimizar riesgos.

3.2. Importancia del pronóstico en la gestión empresarial

El pronóstico de demanda es una herramienta clave para alinear los recursos de la empresa con las necesidades reales del mercado, optimizando costos y mejorando la eficiencia operativa. Según Armstrong (2012), un buen pronóstico permite anticipar cambios en la demanda, reducir riesgos y garantizar la disponibilidad de productos, impactando positivamente en la satisfacción del cliente y la toma de decisiones estratégicas.

3.3. Desafíos en el pronóstico de la demanda en el sector de bebidas

El sector de bebidas, tanto alcohólicas como no alcohólicas, enfrenta desafíos significativos en la predicción de la demanda debido a su volatilidad, estacionalidad y sensibilidad a factores externos como el clima o las festividades. Ramanathan (2012) destaca que una planeación de la demanda adecuada es clave para una cadena de suministro eficiente, mientras que Hübner, Kuhn y Sternbeck (2013) señalan que en países industrializados esta gestión es un factor crítico de éxito. En contraste, en Colombia persisten brechas conceptuales y metodológicas que limitan la integración entre demanda y suministro (GS1, Logyca y Proexport, 2009).

Un pronóstico deficiente puede generar desajustes de inventario, sobrecostos y pérdida de oportunidades. Edenfield (2014) afirma que una planeación bien estructurada proporciona visibilidad en la cadena de valor, anticipando problemas de capacidad o inventario. A nivel operativo, los pronósticos se centran en ventas por SKU y canal de distribución, basados en patrones históricos (Maaß, Spruit & de Waal, 2014). Sin embargo, la incertidumbre y variabilidad del consumo dificultan la precisión, especialmente en productos con ciclos de vida cortos o alta estacionalidad (Christopher, 2004).

Uno de los mayores retos es pronosticar productos nuevos, donde la falta de datos históricos limita el uso de modelos cuantitativos. Por ello, muchas organizaciones recurren a métodos subjetivos como encuestas o juicios expertos, lo que incrementa el margen de error (Goodwin, Meeran y Dyussekeneva, 2014). Además, cerca del 50% de los productos lanzados en consumo masivo fracasan rápidamente por expectativas poco realistas, afectando los pronósticos y decisiones estratégicas (Adams, 2004).

Ante estos retos, se propone la implementación de modelos híbridos que integren métodos tradicionales (Promedio Móvil Simple, Regresión Lineal, Holt-Winters, MSTL con regresión polinómica) con algoritmos de aprendizaje automático como XGBoost y LightGBM, los cuales permiten capturar patrones complejos y mejorar la precisión de los pronósticos en un mercado altamente dinámico.

3.4. Series temporales y modelos tradicionales

El análisis de series temporales es esencial para comprender la evolución de datos en el tiempo, permitiendo detectar tendencias, estacionalidades y fluctuaciones relevantes para la toma de decisiones en áreas como logística, producción y ventas (Hyndman y Athanasopoulos, 2018). Este tipo de análisis descompone los datos en tres componentes clave: tendencia, estacionalidad y variaciones aleatorias, lo cual es especialmente útil en sectores con alta estacionalidad como el de bebidas (Moreno-García et al., 2017).

Las organizaciones utilizan modelos tradicionales como Promedio Móvil Simple, Regresión Lineal y Holt-Winters para capturar patrones básicos en los datos. Sin embargo, cuando las relaciones entre variables son más complejas, se requieren métodos más avanzados, como MSTL con regresión polinómica y algoritmos de aprendizaje automático como XGBoost y LightGBM, que permiten modelar relaciones no lineales y datos dinámicos.

La mejora en la precisión de estos modelos depende de una adecuada optimización de parámetros, la cual puede realizarse mediante técnicas como Grid Search, búsqueda aleatoria u optimización bayesiana (Makridakis et al., 2018). Estas estrategias permiten afinar los modelos y mejorar su capacidad predictiva.

En el sector de bebidas, donde influyen factores como la estacionalidad y la volatilidad del mercado, la correcta aplicación de modelos de series temporales y técnicas de machine learning resulta crucial para optimizar la planeación de la demanda, gestionar inventarios de forma eficiente y fortalecer la competitividad (Christopher, 2004).

3.4.1. Conceptos básicos de series temporales

Para comprender y modelar adecuadamente las series temporales es necesario identificar tres componentes fundamentales: estacionariedad, tendencia y estacionalidad. Estas características permiten descomponer el comportamiento de una variable a lo largo del tiempo y facilitan la selección de modelos predictivos más adecuados (Hamilton, 1994).

- **Estacionariedad:** Es la propiedad de una serie temporal cuyas características estadísticas, como la media y la varianza, se mantienen constantes a lo largo del

tiempo. Una serie no estacionaria suele mostrar tendencias o variaciones estacionales que deben ser tratadas antes de aplicar ciertos modelos predictivos. Una prueba común para validar la estacionariedad es la prueba de Dickey-Fuller aumentada (ADF) (Hyndman & Athanasopoulos, 2018).

- **Tendencia:** Es el movimiento a largo plazo de una serie temporal, ya sea creciente o decreciente. Por ejemplo, en sectores como el de bebidas, la tendencia puede reflejar un crecimiento sostenido de las ventas debido a la expansión de mercados o cambios en las preferencias del consumidor (Makridakis et al., 2008).

- **Estacionalidad:** Representa patrones cíclicos que se repiten a intervalos regulares, como picos de demanda en festividades o temporadas climáticas específicas (Quintana & Jiménez, 2016). En el sector de bebidas, las ventas suelen aumentar en verano o durante celebraciones navideñas.

Estos conceptos son cruciales al modelar la demanda con técnicas como Promedios Móviles Simples, Regresión Lineal, Holt-Winters o enfoques más avanzados como MSTL con regresión polinómica, XGBoost y LightGBM, que requieren una comprensión previa de la estructura de la serie temporal para lograr una mayor precisión en los pronósticos (Hyndman & Athanasopoulos, 2018).

3.4.2. Modelo Holt-Winters

El modelo Holt-Winters es una técnica de suavizado exponencial que permite modelar series de tiempo que presentan tanto tendencia como estacionalidad. Se trata de un método ampliamente empleado para realizar pronósticos de corto y mediano plazo en contextos donde los datos presentan variaciones estacionales regulares y una evolución tendencial (Winters, 1960). Este procedimiento ajusta de forma recursiva tres componentes fundamentales de la serie: el nivel, la tendencia y la estacionalidad, utilizando para ello coeficientes de suavizamiento que ponderan la importancia relativa de las observaciones más recientes frente a las pasadas. El método ofrece dos versiones principales:

- Aditiva, recomendable cuando las fluctuaciones estacionales son constantes a lo largo del tiempo.

- Multiplicativa, más adecuada cuando la amplitud de la estacionalidad varía proporcionalmente al nivel de la serie (Coshall, 2006).

Formulación del modelo Holt-Winters Aditivo:

- Nivel: $l_{t=a(Y_t-8_{t-m})+(1-a)(l_{t-1}+b_{t-1})}$
- Tendencia: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$
- Estacionalidad: $8_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)8_{t-m}$
- Pronóstico para t + h: $\hat{Y}_{t+h} = l_t + hb_t + 8_{t-m+h}$

Formulación del modelo Holt-Winters Multiplicativo:

- Nivel: $l_t = a(Y_t - 8_{t-m}) + (1-a)(l_{t-1} + b_{t-1})$
- Tendencia: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$
- Estacionalidad: $8_t = \gamma(Y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$
- Pronóstico para t + h: $\hat{Y}_{t+h} = (l_t + hb_t)s_{t-m+h}$

Dónde:

- l_t es el nivel,
- b_t es la tendencia,
- s_t es el componente estacional,
- $\hat{Y}_{t+h}$ es el pronóstico h pasos adelante,
- α, β y γ son los parámetros de suavizamiento, todos comprendidos entre 0 y 1,
- m es la longitud del ciclo estacional.

Este enfoque es especialmente útil en sectores como la industria de bebidas y alimentos, donde las ventas suelen estar sujetas a ciclos estacionales claros, por ejemplo, incrementos en épocas festivas o durante el verano (Makridakis et al., 2008). Su popularidad también radica en su capacidad para mitigar el impacto del componente aleatorio o ruido en las series temporales, generando predicciones más estables y confiables (Lim & McAleer, 2001).

A través de la ponderación diferencial de los datos más recientes, el modelo Holt-Winters permite capturar de manera efectiva la dinámica cambiante de la demanda en mercados

volátiles, contribuyendo a una mejor planificación y toma de decisiones estratégicas (Hyndman & Athanasopoulos, 2018).

3.4.3. Regresión lineal simple

La regresión lineal simple es uno de los métodos más básicos y utilizados en análisis predictivo. Su objetivo es modelar la relación entre una variable independiente X (por ejemplo, el tiempo) y una variable dependiente Y (por ejemplo, las ventas), ajustando una línea recta que minimiza la suma de los errores cuadrados entre los valores observados y los valores predichos (Montgomery, Peck & Vining, 2012).

Esta técnica es ampliamente empleada en series temporales para detectar tendencias lineales a lo largo del tiempo. A pesar de su simplicidad, es una herramienta útil para establecer proyecciones iniciales cuando las series no presentan estacionalidad significativa o cuando se desea identificar la dirección general del comportamiento de los datos (Wooldridge, 2016).

Fórmula de la regresión lineal simple: $\hat{Y}_t = \beta_0 + \beta_1 X_t + \varepsilon_t$

Dónde:

- $\hat{Y}_t$: valor estimado de la variable dependiente en el tiempo t
- β_0: Intercepto de la recta
- β_1: Pendiente que mide la tasa de cambio
- X_t: Variable independiente (en series temporales suele ser el tiempo)
- ε_t: Error aleatorio

Características principales:

- Facilidad de implementación
- Ideal para la detección de tendencias lineales en los datos.
- No captura estacionalidad ni patrones no lineales (Hyndman & Athanasopoulos, 2018).

3.4.4. Promedio móvil simple

El Promedio Móvil Simple (SMA, por sus siglas en inglés) es uno de los métodos más básicos y utilizados en el análisis de series temporales. Consiste en calcular la media aritmética de un conjunto de observaciones pasadas, suavizando la serie temporal para resaltar su tendencia subyacente y eliminar parte del "ruido" aleatorio (Makridakis et al., 2008). Este método es útil cuando se desea identificar patrones generales o ciclos de corto plazo.

El SMA es especialmente adecuado para datos sin fuerte estacionalidad ni tendencia pronunciada, aunque también es empleado como referencia inicial antes de aplicar modelos más complejos (Hyndman & Athanasopoulos, 2018).

Formulación del promedio móvil simple:

$$\hat{y}_t = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i}$$

Dónde:

- $\hat{Y}_t$ es el valor suavizado en el tiempo t
- K es la cantidad de períodos considerados en la ventana de promedio
- Y_{t-i} son los valores reales de la serie en los κ períodos anteriores

El SMA es ampliamente utilizado en logística, inventarios y ventas, ya que proporciona una visión clara de la tendencia a corto plazo (Silver et al., 2017).

3.4.5. MSTL con regresión polinómica

El método MSTL (Multiple Seasonal-Trend decomposition using Loess) es una técnica avanzada de descomposición de series temporales que permite extraer múltiples componentes estacionales junto con la tendencia y el residuo de una serie. Esta técnica es

útil cuando los datos presentan más de un ciclo estacional (por ejemplo, anual y semanal) (Bandara et al., 2020).

Una vez descompuesta la serie con MSTL, es posible aplicar una regresión polinómica sobre la componente de tendencia, lo que permite capturar relaciones no lineales en los datos. La regresión polinómica modela la variable dependiente como una función polinómica de la variable independiente (Hyndman & Athanasopoulos, 2018), siendo útil para detectar curvaturas suaves o complejas en la tendencia.

Formula: $Y = \beta_0 + \beta_{1x} + \beta_{2x^2} + \cdots + \beta_{nx^n} + \epsilon$

Dónde:

- Y es la tendencia ajustada
- x es el tiempo (o cualquier variable independiente)
- β_i son los coeficientes del modelo
- ϵ es el térmico de error

3.4.6. XGBoost

El algoritmo XGBoost (Extreme Gradient Boosting) es una técnica de aprendizaje automático supervisado que ha ganado gran popularidad por su eficiencia y alto rendimiento, especialmente en tareas de clasificación y regresión (Chen & Guestrin, 2016). XGBoost es una evolución del algoritmo Gradient Boosting (GBoost), pero introduce mejoras significativas en términos de velocidad y precisión gracias a su capacidad para paralelizar la construcción de árboles de decisión y optimizar el uso de los recursos computacionales (The Machine Learners, 2021).

A diferencia de otros métodos de boosting que construyen árboles de manera secuencial, XGBoost realiza una exploración más eficiente del espacio de búsqueda utilizando técnicas como la poda automática de ramas y la optimización de la función de pérdida a través del segundo orden (gradiente y Hessiano) (XGBoost Developers, 2021). Además, pondera las hojas más puras y realiza una asignación más adecuada de los pesos, mejorando la precisión del modelo.

Características clave de XGBoost en series temporales:

- Capacidad para manejar series temporales con estacionalidad y tendencia mediante la inclusión de variables rezagadas (lags) y variables exógenas.
- Construcción de árboles de decisión adaptativos con una optimización eficiente del hardware.
- Reducción tanto del sesgo como de la varianza en las predicciones, lo que lo hace ideal para conjuntos de datos complejos.

Fórmula simplificada del modelo XGBoost:

$$\hat{Y}_i = \sum_{k=1}^{k} f_K(x_i), \quad f_k \, \epsilon F$$

Dónde:

- f_k representa un árbol de decisión individual
- F es el espacio de funciones de regresores CART
- k es el número total de arboles construidos secuencialmente
- $\hat{Y}_i$ es la predicción final para la observación i

El modelo optimiza la siguiente función objetivo, que combina la función de pérdida (L\mathcal{L}L) con un término de regularización (Ω\OmegaΩ) que controla la complejidad del modelo:

$$Obj = \sum_{i=1}^{n} L(Y_i, \hat{Y}_i) + \sum_{k=1}^{k} \Omega(f_k)$$

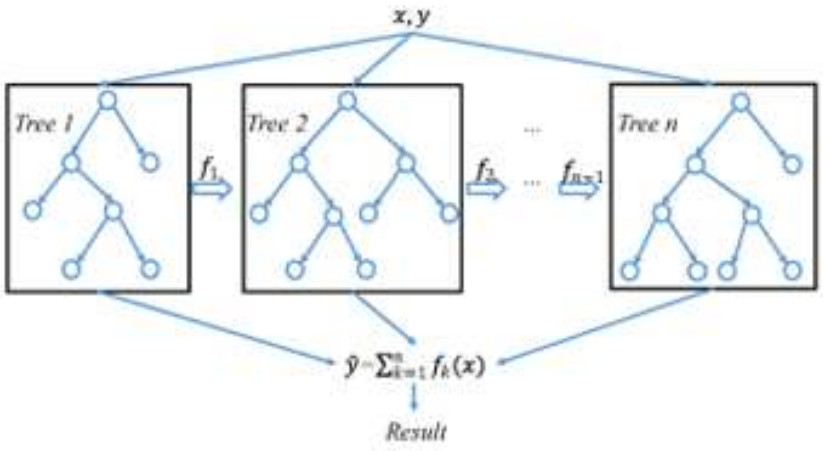

Figura 1. Esquema de XGBoost.

Fuente: ResearchGate, 2019.

3.4.7. LightGBM

LightGBM es una técnica avanzada de aprendizaje automático basada en el algoritmo de potenciación de gradiente, diseñada para ser altamente eficiente y rápida. Desarrollado por Microsoft, LightGBM destaca por su capacidad de manejar grandes volúmenes de datos con bajo consumo de memoria gracias a la construcción de árboles tipo leaf-wise, en lugar del tradicional level-wise (Ke et al., 2017).

Otra de sus ventajas es el uso de histogramas para discretizar los datos continuos, acelerando el proceso de entrenamiento y reduciendo la complejidad computacional (Zhang & Yang, 2020). Además, es especialmente útil para problemas de predicción con grandes conjuntos de datos y múltiples variables, como ocurre frecuentemente en el análisis de series temporales (Chen et al., 2020).

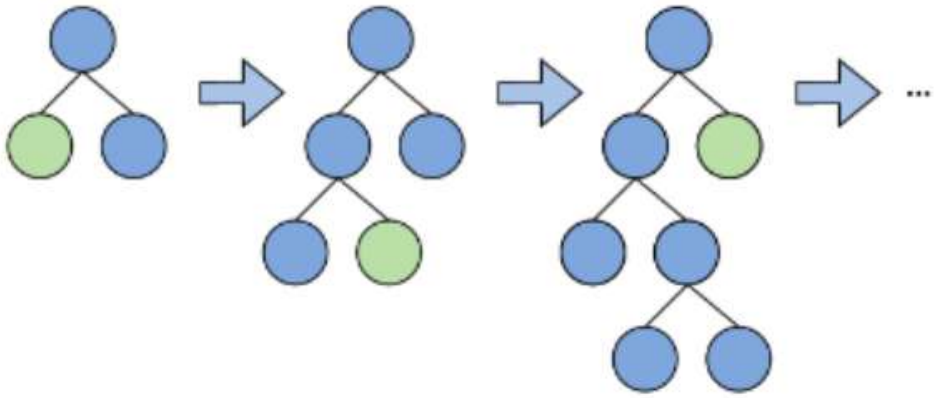

Figura 2.Algoritmo Boosting de machine learning.

Fuente: Medium

LightGBM es capaz de capturar relaciones no lineales y patrones estacionales complejos al integrar variables adicionales (lags, mes, día) dentro de su estructura de árboles de decisión. Esto lo convierte en una herramienta poderosa y versátil en la modelación de series de tiempo y en problemas de forecasting.

Principales ventajas en series temporales:

- Eficiencia y velocidad: LightGBM es más rápido en entrenamiento y predicción en comparación con otros algoritmos de boosting, especialmente en conjuntos de datos grandes.
- Manejo de grandes volúmenes de datos: Está diseñado para manejar grandes conjuntos de datos con menor uso de memoria, lo que es beneficioso en series temporales extensas.
- Capacidad para capturar relaciones no lineales: LightGBM puede modelar relaciones complejas en los datos, lo que es útil para capturar patrones en series temporales.

Formulación: LightGBM sigue el principio del boosting de gradiente, donde se combinan múltiples modelos débiles (generalmente árboles de decisión) para formar un modelo fuerte. Matemáticamente, el modelo se representa como una suma de funciones:

$$\hat{Y}_i = \sum_{\kappa=1}^{K} f_k(X_i), \;\; f_k \;\epsilon\, F$$

Dónde:

$\hat{Y}_i$ es la predicción para la instancia i.

K es el número total de árboles.

f_k representa un árbol de decisión individual perteneciente al espacio de funciones f.

La función objetivo que LightGBM optimiza combina una función de pérdida que evalúa la diferencia entre las predicciones y los valores reales, y un término de regularización que controla la complejidad del modelo:

obj= $\sum_{i=1}^{n} L(Y_i, \hat{Y}_i) + \sum_{K=1}^{k} \Omega(f_k)$

LightGBM demostró ser una herramienta eficiente para la predicción de series temporales, destacándose por su rapidez y bajo consumo de memoria. Su capacidad de manejar grandes volúmenes de datos y su enfoque "leaf-wise" le permitió modelar adecuadamente la tendencia y la estacionalidad de la serie simulada. Aunque es un modelo originalmente diseñado para tareas de clasificación y regresión en tabulares, su flexibilidad al incorporar variables como lags y características estacionales lo convierte en una opción atractiva para series temporales con alta variabilidad y patrones complejos.

3.4.8. Comparativa de modelos

Con el fin de seleccionar el modelo de pronóstico más adecuado para la demanda de Dislicores, es necesario comparar y analizar las principales características de los enfoques implementados. Cada modelo tiene ventajas específicas que lo hacen más o menos apropiado según la naturaleza de los datos y las condiciones del negocio.

A continuación, se presenta una tabla comparativa que resume las principales características de los modelos empleados: Holt-Winters, Regresión Lineal Simple, Promedio Móvil Simple, MSTL con Regresión Polinómica, XGBoost y LightGBM. Esta tabla

permite visualizar de manera clara las diferencias entre modelos tradicionales, de descomposición y de aprendizaje automático, destacando sus aplicaciones más frecuentes en series temporales.

Tabla 1.Comparativa de modelos de pronósticos:

Modelo	Tipo	Características principales	Ventajas	Aplicación en series temporales
Holt-Winters	Suavizado exponencial	Capta tendencia y estacionalidad (aditivo o multiplicativo).	Fácil implementación, útil en datos con estacionalidad clara.	Previsión de ventas estacionales, logística y cadenas de suministro.
Regresión Lineal Simple	Estadístico tradicional	Relación lineal entre la variable dependiente y el tiempo.	Fácil de interpretar y rápida de aplicar.	Series con tendencia lineal sin estacionalidad o con estacionalidad controlada.
Promedio Móvil Simple (SMA)	Suavizado	Suaviza la serie calculando el promedio de los últimos n períodos.	Reduce el ruido y resalta la tendencia general.	Buen preprocesador para eliminar variaciones aleatorias antes de modelar.
MSTL + Regresión Polinómica	Descomposición + Regresión	Descompone múltiples estacionalidades y ajusta la	Capta estacionalidades múltiples y	Series complejas con múltiples ciclos estacionales

		tendencia con regresión polinómica.	tendencias no lineales.	(e.g., mensual + semanal).
XGBoost	Machine Learning (Ensemble)	Algoritmo de gradient boosting eficiente y preciso basado en árboles.	Maneja grandes volúmenes de datos, capta no linealidades y outliers.	Predicción de series temporales multivariantes y contextos de alta complejidad.
LightGBM	Machine Learning (Ensemble)	Usa histogramas y crecimiento "leaf-wise" para construir árboles más eficientes.	Optimiza velocidad y uso de memoria, ideal para grandes datasets.	Series de tiempo con grandes volúmenes de datos y relaciones complejas.

Fuente: Elaboración propia

Recomendación:

- Para Dislicores y el sector de bebidas en general, donde la demanda es altamente estacional y está influenciada por factores externos y de comportamiento del consumidor, se recomienda priorizar XGBoost o LightGBM como modelos principales para la toma de decisiones estratégicas y operativas. No obstante, modelos como Holt-Winters o MSTL pueden complementar la estrategia para pronósticos de corto plazo o categorías con patrones de estacionalidad bien definidos.

3.5. Métricas de evaluación para los pronósticos de los modelos.

Las métricas de evaluación permiten medir la precisión y el desempeño de los modelos de pronóstico, comparando los valores predichos con los valores reales observados. Estas métricas son fundamentales para determinar qué modelo ofrece mejores resultados y tomar decisiones informadas sobre cuál utilizar en la práctica (Hyndman & Athanasopoulos, 2018).

En el caso de series de tiempo, es recomendable emplear métricas que permitan interpretar fácilmente los errores relativos al tamaño de la demanda. Algunas de las más utilizadas son:

- MAE% (Error Absoluto Medio Porcentual): Calcula el error absoluto promedio expresado como porcentaje, permitiendo evaluar la magnitud promedio de los errores sin considerar su dirección. Es útil para comparar errores relativos en diferentes escalas (Makridakis et al., 2008).
- SESGO% (Porcentaje de Sesgo): Mide si el modelo tiende a sobreestimar o subestimar sistemáticamente los valores reales. Un sesgo positivo indica sobreestimación y un sesgo negativo, subestimación (Hyndman & Athanasopoulos, 2018).
- SCORE%: Es una métrica complementaria que combina la precisión y la estabilidad del modelo, proporcionando una visión más completa de su desempeño. Se usa principalmente para ajustar modelos en contextos de negocios y logística (Mentzer & Moon, 2005).

El uso conjunto de estas métricas facilita la selección del modelo más adecuado para contextos reales, como la gestión de inventarios y la planificación de la demanda.

3.5.1. MAE% (Error Absoluto Medio Porcentual)

El MAE% es una métrica que mide el promedio de los errores absolutos entre las predicciones y los valores reales, expresándolo como un porcentaje respecto al valor observado. Es útil para evaluar la precisión del modelo y permite interpretar fácilmente la magnitud del error en términos relativos (Hyndman & Athanasopoulos, 2018).

Fórmula:

$$MAE\% = \left(\frac{1}{n}\sum_{t=1}^{n}\frac{\left|Y_t - \hat{Y}_t\right|}{Y_t}\right) \times 100$$

Dónde:

- Y_t: Valor real en el tiempo t
- $\hat{Y}_t$: Valor pronosticado en el tiempo t
- n: número total de observaciones

El MAE% suma el valor absoluto de los errores relativos en cada período y luego promedia ese valor para obtener un porcentaje global del error respecto a los valores reales.

3.5.2. SESGO% (Porcentaje de Sesgo)

El SESGO% es una métrica que evalúa si un modelo tiende sistemáticamente a sobreestimar o subestimar los valores reales. Expresa el promedio de los errores como porcentaje, ayudando a identificar la dirección y magnitud del sesgo en los pronósticos (Hyndman & Athanasopoulos, 2018).

Fórmula:

$$SESGO\% = \left(\frac{1}{n}\sum_{t=1}^{n}\frac{(Y_t - \hat{Y}_t}{Y_t}\right) \times 100$$

Dónde:

- Y_t: Valor real en el tiempo t

- $\hat{Y}_t$: Valor pronosticado en el tiempo t
- n: número total de observaciones

A diferencia del MAE%, el SESGO% no toma el valor absoluto de los errores, permitiendo así detectar si, en promedio, el modelo tiende a quedarse por encima (sesgo positivo) o por debajo (sesgo negativo) de los valores reales.

3.5.3. SCORE%

El SCORE% es una métrica que mide la precisión de un modelo de pronóstico comparando el error absoluto medio (MAE) del modelo frente a un modelo de referencia o "modelo base" (por ejemplo, un modelo de promedio móvil o naïve). Expresa la mejora relativa del modelo sobre la referencia en términos porcentuales (Hyndman & Athanasopoulos, 2018).

Fórmula:

$$SCORE\% = \left(1 - \frac{\sum_{t=1}^{n} |y_t - y_t|}{\sum_{t=1}^{n} y_t}\right) \times 100$$

Donde:

- y_t Es el valor real en el tiempo t
- $\hat{y}_t$ Es el pronosticado en el tiempo t
- n Es el número total de observaciones

Interpretación: Makridakis, Wheelwright y Hyndman (2008) señalan que:

- SCORE% cercano a 100%: Alta precisión del modelo, con predicciones muy cercanas a los valores reales.

SCORE% bajo: Baja precisión del modelo, lo que indica que los errores son significativos en relación con los valores reales.

3.5.4. RMSE% (Raíz del error cuadrático medio)

El RMSE% es una métrica de evaluación utilizada para cuantificar la magnitud promedio de los errores de un modelo de predicción, penalizando fuertemente los errores grandes. A diferencia del MAE%, el RMSE% eleva al cuadrado las diferencias entre los valores reales y los valores pronosticados antes de promediar, lo que lo hace más sensible a los valores atípicos (Makridakis et al., 2008).

Fórmula:

$$RMSE\% = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i - Y_i}{Y_i}\right)^2} \times 100$$

Donde:

- $\hat{Y}_i$ es el valor pronosticado,
- Y_i es el valor real observado,
- n es el número total de observaciones.

3.6. Revisión de la literatura

En este apartado se presenta una muestra de los principales estudios y teorías que sustentan la aplicación de modelos de pronóstico en series temporales y aprendizaje automático, así como su relevancia en el sector de bebidas y la importancia de las métricas de evaluación.

1. Modelos tradicionales de series temporales:
    - Holt y Winters (1957): Desarrollaron el modelo Holt-Winters, una técnica basada en el suavizado exponencial que permite modelar series con componentes de tendencia y estacionalidad.
    - Regresión Lineal Simple (Montgomery et al., 2012): Este modelo es uno de los métodos más sencillos para pronosticar series temporales, especialmente cuando la relación entre la variable dependiente (ventas, demanda, etc.) y el tiempo es lineal.

- Promedio Móvil Simple (Makridakis et al., 2008): Consiste en suavizar la serie temporal calculando el promedio de un número fijo de observaciones anteriores.

2. Técnicas Avanzadas de Series Temporales:
   - MSTL con Regresión Polinómica (Hyndman & Athanasopoulos, 2018): La descomposición MSTL (Multiple Seasonal-Trend Decomposition using Loess) permite extraer múltiples patrones estacionales de una serie temporal, combinada con la regresión polinómica.

3. Técnicas de Aprendizaje Automático:
   - Chen y Guestrin (2016): Presentan XGBoost, un algoritmo de gradient boosting que ha demostrado ser altamente efectivo en la predicción de series temporales multivariantes. Su capacidad de manejar grandes volúmenes de datos y detectar interacciones complejas entre variables lo hacen ideal para contextos dinámicos como el sector de bebidas.
   - Ke et al. (2017): Proponen LightGBM, una alternativa a XGBoost que optimiza la velocidad y el uso de memoria. Su enfoque leaf-wise para la construcción de árboles de decisión lo convierte en una opción eficiente para el modelado de series temporales con grandes conjuntos de datos, mejorando la precisión y reduciendo tiempos de entrenamiento.

4. Métricas de Evaluación:
   - Hyndman y Athanasopoulos (2018): Señalan que, en el contexto de series temporales, es fundamental emplear métricas porcentuales como el MAE% y RMSE% para interpretar correctamente la precisión de los modelos, especialmente cuando las series tienen escalas cambiantes o alta estacionalidad.
   - Armstrong (2001): Resalta la importancia del SESGO% como métrica clave para detectar si un modelo tiende a sobrestimar o subestimar

la demanda. Esta información es crucial para ajustar los modelos y evitar decisiones incorrectas en la gestión de inventarios.

5. Aplicaciones en el Sector de Bebidas:
    - Gupta y George (2016): Analizan la demanda de bebidas y señalan que esta está altamente influenciada por factores estacionales, eventos promocionales y condiciones climáticas. En este contexto, modelos como Holt-Winters y MSTL son especialmente útiles para capturar estos patrones.

Hübner, Kuhn y Sternbeck (2013): Enfatizan la importancia de integrar modelos avanzados de pronóstico en la cadena de suministro de empresas de consumo masivo, como las del sector de bebidas. La implementación de modelos como LightGBM y XGBoost puede optimizar la gestión de inventarios y distribución, reduciendo costos y mejorando la capacidad de respuesta frente a cambios en la demanda.

4. Metodología implementada

El enfoque metodológico adoptado en este proyecto combina técnicas cualitativas y cuantitativas para abordar de manera integral los desafíos de Dislicores en la predicción de su demanda de productos. Este enfoque mixto permite no solo analizar los datos históricos de ventas, sino también interpretar las dinámicas de negocio, el comportamiento del consumidor y los factores exógenos que influyen en la variabilidad de la demanda en el sector de bebidas.

La investigación se desarrolla bajo el marco de la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), reconocida por su estructura flexible y por ser uno de los estándares más utilizados en proyectos de minería de datos y analítica avanzada (Wirth & Hipp, 2000). Su naturaleza iterativa permite revisar y mejorar continuamente cada etapa del proceso, garantizando una alineación constante con los objetivos de negocio y la calidad de los datos disponibles.

La implementación de CRISP-DM en este proyecto asegura una transición fluida desde la comprensión del negocio y la exploración de datos hasta la construcción, evaluación y despliegue de modelos predictivos. Esto resulta clave para asegurar la integración exitosa

de los modelos en los procesos operativos de Dislicores, mejorando la precisión de las proyecciones y la toma de decisiones estratégicas.

4.1. Diseño de la investigación

El diseño de la investigación adopta un enfoque mixto, combinando técnicas cualitativas y cuantitativas para abordar de manera integral la problemática de Dislicores respecto a la precisión de sus pronósticos de ventas. Esta combinación busca fortalecer la comprensión tanto del contexto empresarial como del comportamiento de la demanda reflejado en los datos históricos.

El marco metodológico que estructura este estudio es CRISP-DM (Cross-Industry Standard Process for Data Mining), ampliamente reconocido por su flexibilidad y eficacia en proyectos complejos de análisis de datos y minería de información (Wirth & Hipp, 2000). Esta metodología proporciona una hoja de ruta para la exploración de datos, la construcción de modelos y la implementación de soluciones predictivas que puedan integrarse de manera efectiva en la toma de decisiones de la empresa.

4.1.1. Enfoque cualitativo y cuantitativo

1. Enfoque cualitativo:

    a. Propósito: Comprender el entorno de negocio de Dislicores, incluyendo las expectativas de los responsables de la organización, las características del mercado y todos aquellos factores exógenes que pueden afectar la dinámica de la demanda, ejemplo: Comportamiento del consumidor, campañas, eventos, etc.

    b. Método: Se realizaron reuniones semiestructuradas con gerentes y directores de ventas y logística para indagar sobre el comportamiento de la demanda, donde se habló sobre los desafíos, limitaciones, y acciones hechas por dichos equipos. Este enfoque permite analizar y enriquecer la interpretación de los resultados cuantitativos. Como menciona Creswell

(2014), la perspectiva cualitativa es clave para obtener una comprensión más profunda de las dinámicas internas y externas de la organización.

2. Enfoque cuantitativo:

    a. Propósito: Con el fin de la construcción de modelos inferenciales que tengas una calidad de pronósticos aceptables, precisos y eficientes, se realizó análisis y modelamiento de datos de la demanda histórica de Dislicores, esto mediante técnicas estadísticas y aprendizaje automático.

    b. Se procesaron datos históricos de ventas (2020-2024), así como información relevante como promociones, campañas y eventos. Los datos obtenidos se sometieron a técnicas de modelado, utilizando algoritmos como: Holt-Winters, Regresión Lineal, MSTL con Regresión Polinómica, XGBoost, LightGBM y Promedios Móviles Simples. La evaluación de los modelos es realizada mediante métricas como: MAE%, RMSE% y SESGO%, siguiendo las recomendaciones de Montgomery, Jennings y Kulahci (2015) para validar patrones y minimizar errores.

4.1.2. Justificación del uso de la metodología CRISP-DM

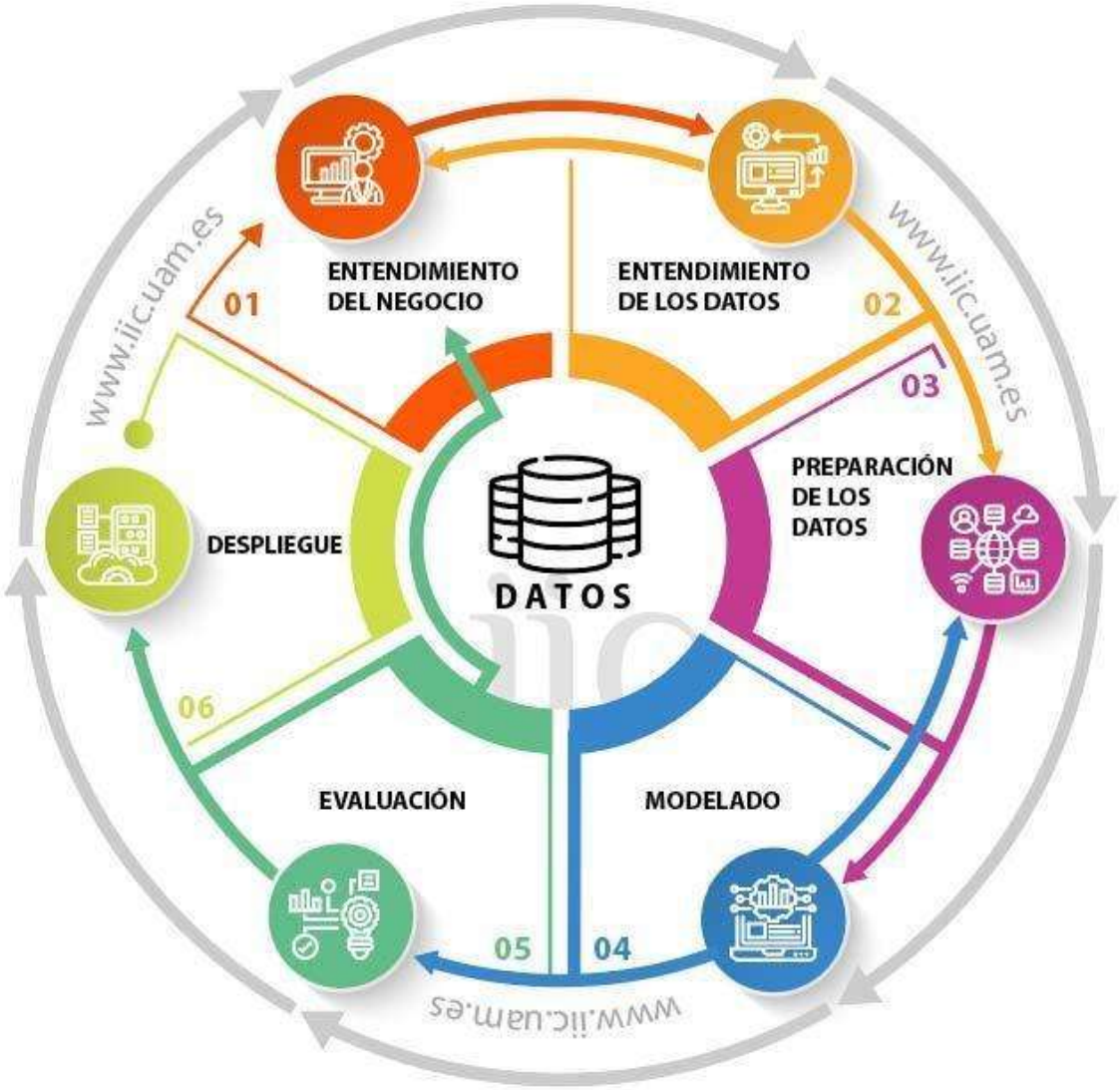

Figura 3.Metodología CRISP-DM y sus fases.

Fuente: Instituto de Ingeniería del Conocimiento (IIC), 2023.

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es el estándar más utilizado en proyectos de análisis de datos debido a su adaptabilidad y enfoque estructurado (Wirth & Hipp, 2000). Su adopción en esta investigación está justificada por:

- Flexibilidad: Según Chapman et al. (2000), CRISP-DM es adaptable a diversos problemas empresariales y permite iterar en las fases para ajustar los modelos y objetivos a los cambios en las necesidades del negocio.

- Estructura clara: La división en seis fases asegura una comprensión profunda del negocio, un análisis riguroso de los datos y una implementación práctica de los modelos (Hofmann & Klinkenberg, 2013).

- Resultados accionables: Su enfoque en la implementación de modelos garantiza que los resultados sean aplicables en la planificación estratégica de Dislicores, optimizando decisiones relacionadas con inventarios, promociones y logística.

- Éxitos previos: Estudios como el de Hyndman y Athanasopoulos (2018) han demostrado que CRISP-DM es eficaz para integrar modelos estadísticos y de aprendizaje automático en proyectos de pronóstico de demanda.

Esta metodología es particularmente adecuada para Dislicores, dado que el proyecto requiere no solo la creación de modelos de pronóstico, sino también su integración en procesos operativos y de toma de decisiones.

4.2. Fases de la metodología CRISP-DM

La metodología CRISP-DM está compuesta por seis fases interconectadas que proporcionan una guía estructurada para abordar proyectos de análisis de datos. Estas fases incluyen la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y el despliegue (Wirth & Hipp, 2000). En esta sección, se aborda la primera fase con un enfoque en las necesidades y objetivos de Dislicores, enriquecida con ejemplos prácticos y casos de éxito.

Las fases de CRISP-DM comprenden:

1. **Comprensión del negocio:** Identificación de los objetivos comerciales de Dislicores.

2. **Comprensión de los datos:** Análisis exploratorio de datos sobre la demanda histórica, así como de otras dimensiones para evaluar la calidad y relevancia de dichos datos.

3. **Preparación de los datos:** Procesos de limpieza y transformación de los datos para la disponibilidad en la implementación de los modelos.

4. **Modelado:** Desarrollar y entrenar los modelos con la demanda histórica procesada.

5. **Evaluación:** Comparar el resultado de los modelos mediante las métricas ya establecidas.

6. **Despliegue:** Implementar una arquitectura de solución con integración de la plataforma de business Intelligence de Dislicores.

4.2.1. Fase 1: Comprensión del negocio

Esta fase busca alinear las soluciones analíticas con los objetivos estratégicos de Dislicores. abordando sus desafíos operativos en un entorno competitivo y volátil (Chapman et al., 2000). En este contexto, se analizó la situación actual de la empresa y se identificaron las principales necesidades mediante entrevistas con actores clave del negocio.

Dislicores es un referente en la distribución de bebidas alcohólicas y no alcohólicas en Colombia, gestionando un portafolio con más de 1,000 referencias, donde el 20% de los productos genera la mayoría de las ventas (principio de Pareto). La demanda presenta estacionalidad significativa, con picos en diciembre, Semana Santa y durante eventos especiales (Gupta & George, 2016). Además, la demanda es susceptible a factores externos como promociones o eventos deportivos.

Necesidades identificadas:

- Mejorar la precisión de los pronósticos, superando limitaciones de métodos tradicionales como promedios móviles y juicio experto.
- Optimizar la gestión de inventarios para reducir tanto el sobre stock como las roturas.

- Aumentar la eficiencia logística mediante rutas y almacenamiento ajustados a la demanda.

Objetivos estratégicos:

- Incrementar la precisión de los pronósticos de demanda en al menos un 20%.
- Reducir en un 15% los costos asociados a la mala gestión del inventario.
- Optimizar la logística de distribución para reducir costos operativos.
- Garantizar la disponibilidad de productos clave en temporadas de alta demanda, mejorando la satisfacción del cliente (Hyndman & Athanasopoulos, 2018).

Este enfoque permitirá a Dislicores anticiparse mejor a las fluctuaciones del mercado y fortalecer su competitividad.

4.2.2. Fase 2: Comprensión de los datos

La comprensión y preparación de los datos son etapas clave en la metodología CRISP-DM, ya que permiten construir modelos de pronóstico basados en datos relevantes y de calidad (Wirth & Hipp, 2000). En este proyecto, se trabajó con información histórica de Dislicores para el período 2020-2024, incluyendo ventas diarias y mensuales, así como variables externas como promociones y festividades.

- Descripción de la fuente de datos: Se analizó junto al director de cadena de suministro, la selección de 232 referencias claves de más de 1000 productos activos, priorizando dicha selección bajo el principio de Pareto (80/20).

- Limpieza y depuración: En este proyecto, la fuente de datos proviene de un sistema de gestión refinado y bien estructurado (Datalake en AWS). Tras realizar una verificación exhaustiva de la calidad de los datos, se comprobó que no existían valores atípicos ni datos faltantes significativos. Por esta razón, no fue necesario aplicar procesos adicionales de imputación ni eliminación de outliers. Únicamente se procedió a estandarizar las variables externas para garantizar su correcta integración en los modelos de pronóstico (Hyndman & Athanasopoulos, 2018).

- EDA (Análisis Exploratorio de Datos): Se identificaron patrones de tendencia y estacionalidad mediante descomposición MSTL, y se analizaron correlaciones entre

las ventas y variables externas (Montgomery, Jennings & Kulahci, 2015). Esto permitió obtener información clave para la posterior selección de modelos y garantizar la alineación con los objetivos estratégicos de Dislicores.

4.2.3. Fase 3: Preparación de los datos

La preparación de los datos consiste en transformar la información en un formato adecuado para el modelado, asegurando su correcta interpretación por los algoritmos. Según Hyndman y Athanasopoulos (2018), esta etapa es esencial para optimizar la precisión de los modelos de pronóstico.

En este proyecto, la preparación incluyó la creación de variables derivadas, la agregación de datos a nivel mensual y la integración de variables externas relevantes, como promociones y venta de kits. Adicionalmente, se estructuraron las series de tiempo para capturar la estacionalidad y tendencias específicas del negocio (Wirth & Hipp, 2000; Montgomery et al., 2015).

4.2.4. Fase 4: Modelado

El modelado es una fase clave de la metodología CRISP-DM, ya que permite evaluar y comparar distintas técnicas de pronóstico para seleccionar la que mejor se adapte a las necesidades de Dislicores. En este proyecto, se implementaron tanto modelos tradicionales como algoritmos avanzados de aprendizaje automático, buscando capturar la estacionalidad, la tendencia y las fluctuaciones propias de la demanda. Modelos implementados:

1. Modelos implementados:

    - Holt-Winters:
        - Objetivo: Modelar series de tiempo con estacionalidad y tendencia suave.
        - Implementación: Se aplicó el método de suavización exponencial del Holt-Winters, utilizando la librería statsmodels en Python. Los parámetros de nivel, tendencia y estacionalidad se ajustaron mediante optimización automática.

- Regresión lineal simple:
  - Objetivo: Capturar la tendencia general en la demanda.
  - Implementación: Se empleó la librería scikit-learn para ajustar un modelo de regresión lineal, útil para escenarios donde la demanda presenta un comportamiento estable y lineal en el tiempo.

- Promedio Móvil Simple (SMA):
  - Objetivo: Suavizar la serie temporal para capturar la tendencia a corto plazo.
  - Implementación: Se aplicó un enfoque de promedio móvil con distintas ventanas (k) para identificar la mejor configuración, implementado en pandas.

- MSTL con Regresión Polinómica:
  - Objetivo: Descomponer la serie en tendencia, estacionalidad y residuos mediante descomposición estacional (MSTL) y luego ajustar una regresión polinómica sobre la tendencia.
  - Implementación: Se utilizó la librería statsmodels para la descomposición y scikit-learn para la regresión polinómica, permitiendo capturar relaciones no lineales en la tendencia.

- XGBoost:
  - Objetivo: Incorporar variables externas (promociones, clima, eventos) y capturar relaciones complejas en la demanda.
  - Implementación: Se utilizó la librería XGBoost en Python, optimizando hiperparámetro mediante Grid Search. Este modelo fue útil para identificar interacciones no lineales entre la demanda y factores externos.

- LigthGBM:
  - Objetivo: Proveer una alternativa más eficiente y rápida que XGBoost, especialmente para manejar grandes volúmenes de datos.

- Implementación: Se aplicó LightGBM en Python, aprovechando su capacidad para gestionar datos categóricos y construir árboles mediante crecimiento foliar (leaf-wise), lo que mejora la precisión en series complejas.

2. Herramientas usadas:

    - Python (Jupyter Notebook): Desarrollo de los modelos y documentación interactiva.
    - Librerias: statsmodels, scikit-learn, XGBoost, LightGBM, pandas, NumPy, etc.
    - AWS (Amazon Web Services): S3 para almacenamiento y EC2 para procesamiento en la nube.
    - Qlik Sense: Visualización y comunicación de resultados con Dashboard's interactivos.

4.2.5. Fase 5: Evaluación

La fase de evaluación es clave para garantizar que los modelos seleccionados sean precisos y útiles para la operación de Dislicores. Se utilizó una división de los datos históricos en conjuntos de entrenamiento (80%) y prueba (20%), junto con validación cruzada temporal, que respeta la secuencia cronológica y evita fugas de información (Bergmeir & Benitez, 2012).

La comparación de los modelos se realizó utilizando tres métricas específicas:

- MAE% (Error Absoluto Medio Porcentual)
- SESGO% (Porcentaje de Sesgo)
- SCORE%

las cuales son ampliamente recomendadas en el ámbito de las series temporales (Hyndman & Athanasopoulos, 2018). Estas métricas permiten evaluar la precisión general, identificar sesgos sistemáticos y ofrecer una visión completa del rendimiento de los modelos (Montgomery, Jennings & Kulahci, 2015).

El uso combinado de estas técnicas y métricas garantizó la selección del modelo que mejor se adapta a las necesidades de Dislicores, optimizando la predicción de la demanda y reduciendo los costos asociados a inventarios y logística.

4.2.6. Fase 6: Despliegue

El modelo de pronóstico fue desplegado utilizando una arquitectura basada en Amazon Web Services (AWS) para garantizar eficiencia, automatización y escalabilidad. La solución se implementó mediante contenedores Docker ejecutados en AWS Lambda, gestionados a través de Amazon Elastic Container Registry (ECR) y automatizados con EventBridge. Los datos históricos y dimensionales de Dislicores se almacenan en Amazon S3 en formato Parquet, permitiendo una lectura y procesamiento eficiente (Merkel, 2014; Jamsa, 2013).

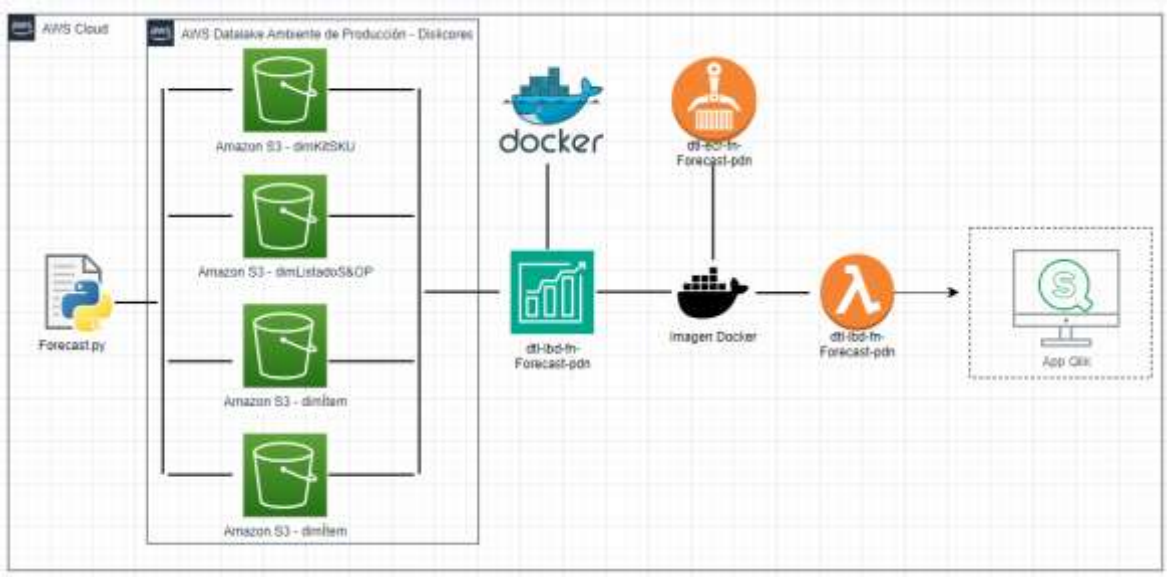

Figura 4.Diagrama de arquitectura del despliegue del modelo de pronóstico.

Fuente: Elaboración propia.

El flujo automatizado garantiza la actualización mensual de las predicciones y su visualización en Dashboard's interactivos de Qlik Sense, facilitando la toma de decisiones operativas y estratégicas (Qlik, 2021). Esta infraestructura permite a Dislicores integrar sus modelos de pronóstico directamente en su ecosistema operativo, mejorando la eficiencia logística y la gestión de inventarios.

Tal como recomiendan Hyndman y Athanasopoulos (2018), la automatización y despliegue continuo son esenciales para optimizar la precisión y utilidad de los pronósticos en contextos dinámicos.

4.3. Consideraciones éticas y legales

El desarrollo e implementación de modelos de pronóstico en Dislicores implica importantes responsabilidades éticas y legales. Estas se centran en garantizar la protección de datos, la transparencia de los procesos y el cumplimiento normativo, fundamentales para la sostenibilidad del proyecto y la confianza de los clientes (Floridi et al., 2018).

1. Protección de la privacidad: La gestión de datos históricos debe cumplir con la Ley 1581 de 2012 en Colombia y el GDPR (Reglamento general de protección de datos). Se priorizó la anonimización de la información y la implementación de medidas de seguridad como cifrado de datos y control de accesos con AWS IAM (Voigt & Von dem Bussche, 2017). Casos como el de Airbnb refuerzan la importancia de aplicar estas buenas prácticas (Miller, 2017).

2. Transparencia del modelo: Para reducir la opacidad de los modelos avanzados (como XGBoost o LightGBM), se documentó exhaustivamente el proceso y se habilitaron Dashboard’s claros en Qlik Sense. Esto facilita la interpretación de resultados y garantiza la auditabilidad del proceso (Ribeiro et al., 2016).

3. Sesgo y equidad: Los modelos fueron revisados para detectar sesgos potenciales derivados de estacionalidades o promociones históricas. Se aplicaron validaciones cruzadas y revisiones periódicas para mitigar estos riesgos, asegurando decisiones comerciales más justas (Hyndman & Athanasopoulos, 2018).

4. Cumplimiento normativo: Además de las normativas internacionales, Dislicores garantiza la conformidad con la legislación local (Ley 1581 de 2012), siguiendo ejemplos de éxito de organizaciones como Grupo Éxito (GS1 Colombia, 2019).

En conclusión, este proyecto integra principios éticos y legales sólidos para proteger la privacidad de los datos, garantizar la transparencia en los modelos y cumplir con la regulación vigente, fortaleciendo la confianza de clientes y aliados estratégicos.

5. Análisis exploratorio de datos (EDA)

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) es una etapa fundamental para comprender el comportamiento y las características de los datos antes de construir los modelos de pronóstico. Según Tukey (1977), el EDA tiene como objetivo identificar patrones, anomalías, relaciones entre variables y características clave que influyen en la variable de interés, en este caso, la demanda de productos en Dislicores. Esta etapa también facilita la preparación adecuada de los datos para mejorar el rendimiento de los modelos predictivos.

Objetivos del análisis exploratorio de datos:

- Comprender la estructura y distribución de los datos.
- Detectar patrones estacionales y de tendencia.
- Identificar anomalías o valores atípicos (outliers).
- Evaluar la correlación entre variables internas (ventas, promociones) y externas (clima, festividades).
- Preparar los datos para las siguientes fases del proyecto, como el modelado y la validación.

5.1. Descripción de los datos

La descripción de los datos es una etapa esencial dentro del análisis exploratorio de datos (EDA), ya que permite obtener una visión clara de la estructura y calidad del conjunto de datos que será utilizado para el proyecto de pronóstico de demanda. Este paso es crucial para asegurar que los datos sean pertinentes y estén listos para el modelado.

- Importancia de describir los datos:
    - Entender la información disponible: Identificar qué variables existen, su propósito y su granularidad (Tukey, 1977).

- Detectar problemas de calidad: Evaluar si hay datos faltantes, atípicos o inconsistentes.
- Seleccionar los modelos adecuados: Determinar si la serie presenta estacionalidad, tendencias o depende de factores externos.
- Validar la cobertura temporal: Verificar si los datos históricos abarcan un rango suficiente para capturar los patrones relevantes.

- Datos utilizados en este proyecto:
  - FactVenta: Datos de ventas históricas, detalladas por producto, fecha y región, clave para identificar patrones de demanda.
  - Dimensión de Kits: Información sobre agrupaciones de productos, útiles para distinguir entre ventas regulares y promociones.
  - Listado S&OP: Relación de productos estratégicos validados por la dirección de la cadena de suministro.
  - Dimensión de Ítems: Descripción detallada de cada producto (categoría, tipo de bebida, presentación).
  - Variables externas: Datos sobre clima y eventos especiales que afectan el consumo (ej. incremento en ventas durante festividades o temporadas cálidas).

Como señala Tukey (1977), una adecuada descripción inicial de los datos orienta el análisis y maximiza el valor que se puede extraer del proceso.

5.1.1. Características generales de los datos de ventas

La tabla FactVentas es la principal fuente de datos utilizada en este proyecto para el análisis y predicción de la demanda en Dislicores. Se trata de una tabla transaccional (Tipo I) que almacena información detallada sobre las ventas tanto en el canal B2C como en el canal B2B.

- Principales características:
  - Ubicación: Datalake de producción de Dislicores (Amazon S3).
  - Frecuencia de actualización: Diaria, lo que garantiza datos actualizados y relevantes para los pronósticos.

- Contenido: Datos de ventas históricas desagregadas por producto, cliente, ubicación y fecha.

- Campos clave:
  - Temporalidad: Campos como nkFecha, dtmFechaDoc, numAño y numMes permiten construir series temporales para detectar patrones de tendencia y estacionalidad.
  - Información de producto y ventas: Incluye nkItem, numCantidadFacturada y numValorVenta, que son cruciales para evaluar la demanda por referencia.
  - Promociones y descuentos: Campos como numValorDescuentoLinea y numValorDescuentoGlobal facilitan el análisis del impacto de promociones.
  - Datos de cliente y ubicación: Variables como nkCliente, nkCentroOperacion y nkBodega permiten realizar segmentaciones por zona geográfica y tipo de cliente.

  .

Tabla 2.Especificación de la tabla FactVenta.

Especificación de la tabla FactVenta					
Nombre DataMart		**Nombre Corto**			
Descripción	Tabla de hechos donde se tendrá almacenada la información de ventas del negocio (B2C y B2B)				
Periodo de Carga	Diaria	**Tipo Carga**	Delete / Insert	**Tipo**	Tipo I
Campo	**Descripción**		**Update?**	**Tipo de dato**	**Default Rechazo**
nkFecha			NO	VARCHAR	
nkBodega			NO	VARCHAR	
nkCentroOperacion			NO	VARCHAR	
nkCliente			NO	VARCHAR	
nkItem			NO	VARCHAR	
nkMotivos			NO	VARCHAR	
strIdCia			NO	VARCHAR	
strTipoNegocio			NO	VARCHAR	

strTipoVenta		NO	VARCHAR	
strSubtipoVenta		NO	VARCHAR	
strTipoEntrega		NO	VARCHAR	
strIdTipoDoc		NO	VARCHAR	
strNumDoc		NO	VARCHAR	
strIdCodPago		NO	VARCHAR	
strUndMedida		NO	VARCHAR	
numCantidadFacturada		NO	DECIMAL	
numCostoPromedioUnd		NO	DECIMAL	
numCostoPromedioTotal		NO	DECIMAL	
numPrecioUnd		NO	DECIMAL	
numValorBruto		NO	DECIMAL	
numValorDescuentoLinea		NO	DECIMAL	
numValorDescuentoGlobal		NO	DECIMAL	
numValorImpuesto		NO	DECIMAL	
numValorNeto		NO	DECIMAL	
numValorIco		NO	DECIMAL	
numValorVenta		NO	DECIMAL	
numAño		NO	SMALLINT	
numMes		NO	SMALLINT	
numHora		NO	SMALLINT	
numMinuto		NO	SMALLINT	
IndObsequio		NO	BOOLEAN	
dtmFechaDoc		NO	DATETIME	
dtmFechaActualizacion		NO	DATETIME	Getdate()

Fuente: Elaboración propia

Esta tabla permite crear modelos de predicción robustos al ofrecer datos granulares que ayudan a comprender la dinámica de la demanda y sus variaciones a lo largo del tiempo.

Como señalan Hyndman y Athanasopoulos (2018), una correcta estructuración y actualización de los datos históricos de ventas es esencial para desarrollar pronósticos precisos y relevantes para la gestión de inventarios y la planificación comercial.

5.1.2. Variables estacionales y exógenas consideradas

La incorporación de variables estacionales y exógenas es clave para mejorar la precisión de los modelos de predicción de demanda en Dislicores, ya que permiten capturar tanto patrones recurrentes como factores externos que afectan las ventas (Hyndman & Athanasopoulos, 2018).

- Variables estacionales:
    - Mes y día de la semana: Reflejan patrones de consumo asociados a festividades y comportamiento habitual de los clientes.
    - Estación del año: Relación entre clima (invierno/verano) y el tipo de bebida más demandada (Montgomery et al., 2015).
    - Eventos promocionales: Impacto positivo de campañas y eventos especiales en los picos de ventas.
- Variables exógenas:
    - Clima: La temperatura y la precipitación afectan la preferencia por bebidas frías o calientes.
    - Factores económicos: Variables como la inflación o el PIB influyen en la capacidad de compra de los consumidores.
    - Gasto publicitario: Aumenta la visibilidad y la atracción de productos, modificando la demanda (Chen & Guestrin, 2016).
- Impacto en los modelos:
    - Modelos como Holt-Winters aprovechan las variables estacionales, mientras que algoritmos avanzados como XGBoost y LightGBM incorporan variables exógenas para capturar relaciones más complejas, mejorando la exactitud del pronóstico y apoyando la toma de decisiones en logística y ventas.

5.2. Validación de los datos

La limpieza y procesamiento de datos son pasos clave para asegurar la calidad de la información que alimenta los modelos de predicción de demanda. Aunque los datos de Dislicores provienen de fuentes validadas, se llevó a cabo una verificación adicional para evitar errores que puedan impactar negativamente los resultados (Hyndman & Athanasopoulos, 2018).

- Actividades realizadas:
    - Eliminación de duplicados: Se identificaron y eliminaron registros repetidos que podrían distorsionar las métricas.
    - Revisión de valores nulos: Se detectaron campos incompletos y se definió su tratamiento según la relevancia de la variable.
    - Estandarización de formatos: Se normalizaron fechas y tipos de datos numéricos, garantizando su correcto uso en los modelos.
    - Detección de valores atípicos: Se revisaron ventas inusualmente altas o bajas, evaluando su validez o necesidad de ajuste (Montgomery et al., 2015).

Este proceso es esencial para asegurar que el modelo de pronóstico trabaje sobre una base de datos limpia y confiable, maximizando la precisión y relevancia de los resultados.

5.2.1. Validación de limpieza de datos para la dimensión SKU

Tras la validación y análisis del conjunto de datos df_dimSku, se obtuvieron los siguientes hallazgos:

- Valores Nulos: No se encontraron valores nulos en ninguna de las columnas, lo que indica que los datos están completos y listos para su análisis sin necesidad de imputación.
- Duplicados: No se identificaron registros duplicados, asegurando que cada fila representa información única.

- Tipos de Datos: Todas las variables tienen el tipo de dato esperado (string para identificadores y nombres, y float64 para valores numéricos), por lo que no se requieren conversiones adicionales.
- Valores Atípicos: Se identificó un valor máximo de 700 en la columna numCantRequerida, lo que representaría una posible anomalía en comparación con la mediana (1.0) y el percentil 75 (2.0). Sin embargo, este valor es válido dentro del contexto del negocio, no requiere ajuste.

El conjunto de datos df_dimSku se encuentra en buen estado y puede ser utilizado en el análisis sin requerir una limpieza adicional.

5.2.2. Validación de limpieza de datos para la dimensión Ítem

Tras la validación y análisis del conjunto de datos df_dimItem, se obtuvieron los siguientes hallazgos:

- Calidad general adecuada: El dataset contiene 232 registros sin duplicados y con una estructura consistente de 4 columnas clave: nkItem, strCodReferencia, strReferencia y numFactorEmpaque.

- Datos completos en casi todas las variables: Se detectó 1 valor nulo únicamente en la columna numFactorEmpaque, mientras que las columnas de identificadores y descripciones (nkItem, strCodReferencia, strReferencia) están completas, garantizando integridad en los datos clave de identificación de ítems.

- Datos consistentes en tipos de variable: Las columnas presentan tipos de datos correctos: string para identificadores y descripciones, y float64 para la métrica numérica numFactorEmpaque.

- Comportamiento esperado de la variable numérica: Las estadísticas descriptivas de numFactorEmpaque muestran un promedio de 10.95, con un rango entre 3 y 48, lo que es coherente con las unidades de empaque típicas del sector bebidas (cajas, botellas o presentaciones multipack).

- Sin problemas de duplicidad: La validación de duplicados arrojó 0 registros repetidos, asegurando que no existen filas redundantes en esta dimensión.

El dataset de la dimensión Ítem se encuentra en condiciones adecuadas para su uso en el modelado y análisis de demanda.

5.2.3. Validación de limpieza de datos para la tabla de hechos FactVenta

El dataset FactVenta cuenta con 19.85 millones de registros y 7 columnas esenciales que contienen datos completos. Su estructura es adecuada para análisis de ventas a nivel transaccional, permitiendo realizar estudios temporales (por año y mes) y económicos.

No se identifican valores nulos en ninguna de las columnas de la tabla FactVenta. Esto garantiza que no será necesario realizar imputaciones o limpieza adicional en esta dimensión para el modelado y análisis de series temporales.

Los tipos de datos son correctos y apropiados para el análisis. Las columnas categóricas y de identificación están en formato string, mientras que las columnas numéricas como numCantidadFacturada, numValorVenta, numAnio y numMes se encuentran en formatos numéricos (float64 e int64). Esto facilita la manipulación y el modelado posterior sin necesidad de transformaciones adicionales.

Los valores atípicos detectados en numCantidadFacturada y numValorVenta, como cantidades negativas y valores de venta negativos, se explican por la presencia de devoluciones y notas crédito aplicadas a las ventas. Este comportamiento es común en operaciones comerciales y será considerado al momento de la construcción de los modelos, ya que estas correcciones suelen regularizarse en los meses siguientes y forman parte de la dinámica normal del negocio.

Se identificaron 198,502 registros con valores de venta superiores al percentil 99, considerados como ventas altas. Estas operaciones corresponden principalmente a ventas mayoristas o pedidos corporativos de grandes volúmenes, lo cual es consistente con la dinámica comercial de Dislicores, que atiende tanto a clientes minoristas como a clientes B2B de gran escala.

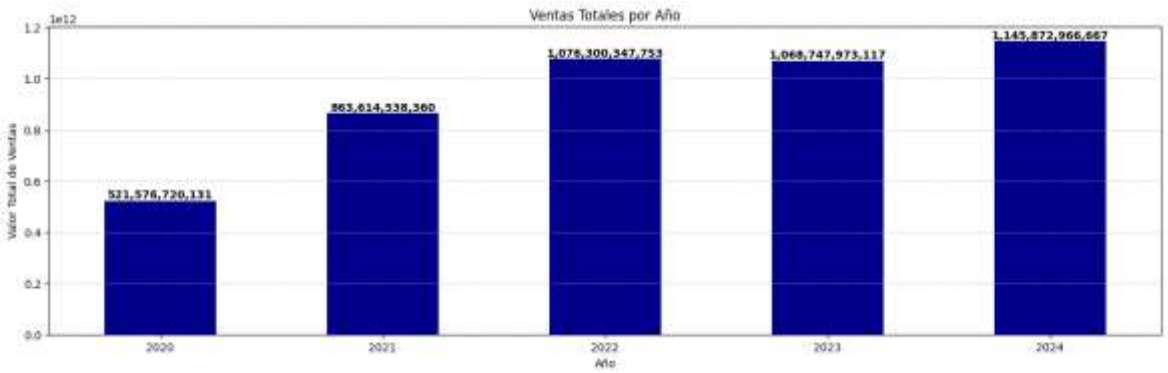

Figura 5.Ventas totales por año.

Fuente: Elaboración propia.

Las ventas totales muestran una tendencia de crecimiento sostenido entre 2020 y 2024. Se observa un fuerte aumento en las ventas desde 2020 hasta 2022, seguido de una leve estabilización en 2023 y un repunte nuevamente en 2024, alcanzando su valor más alto. Esto es consistente con la recuperación postpandemia y la consolidación de Dislicores en el mercado.

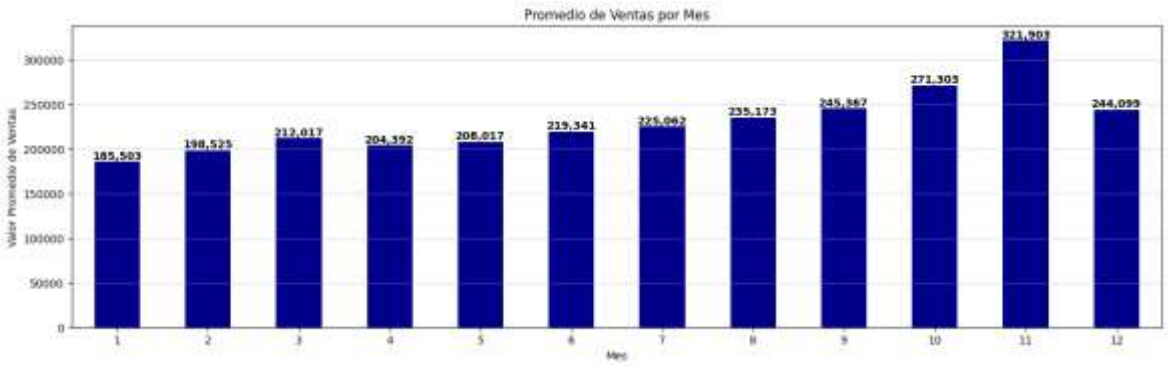

Figura 6. Ventas promedio por mes.

Fuente: Elaboración propia.

El análisis muestra una clara estacionalidad en las ventas, destacándose un incremento progresivo a partir de mitad de año, con picos significativos en octubre y noviembre. Esto puede deberse a eventos de alto consumo como promociones de fin de año y festividades.

Además, enero y febrero presentan los promedios de ventas más bajos, lo que es típico tras la alta demanda de la temporada navideña.

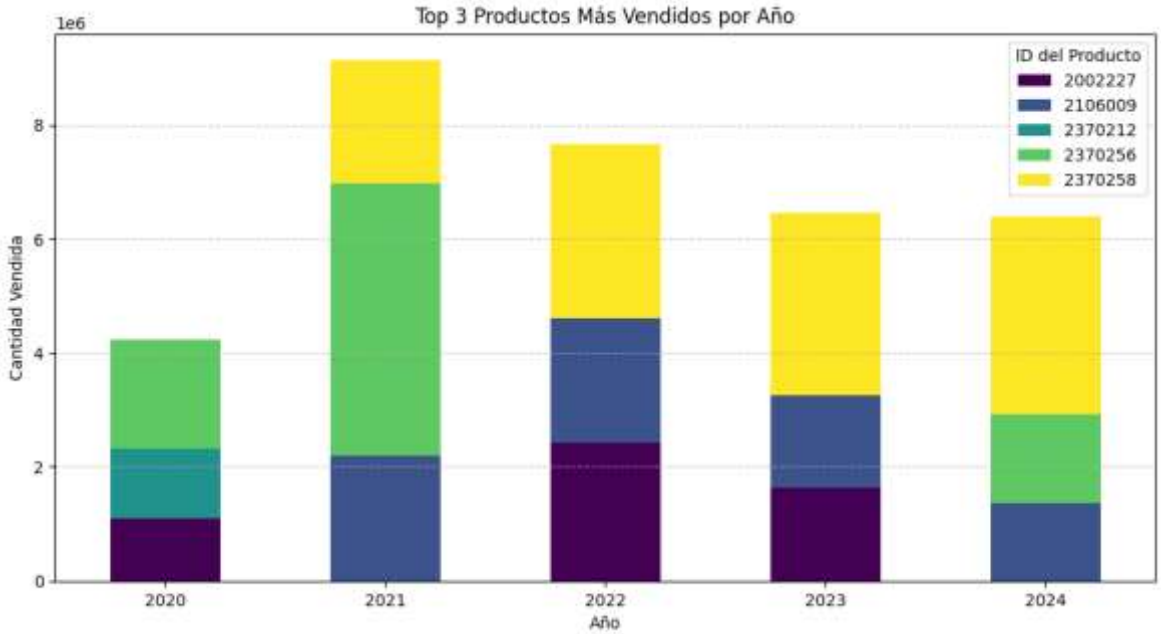

Figura 7.Top 3 productos con más ventas por año.

Fuente: Elaboración propia.

El análisis muestra que los productos top cambian entre años, aunque algunos artículos como el ID 2370258 y 2370256 mantienen una fuerte presencia en varios períodos. Se observa una concentración de las ventas en un pequeño grupo de referencias clave, lo que valida la aplicación del principio de Pareto (80/20) en Dislicores. Esta información es crucial para priorizar la planificación de la demanda y la gestión de inventarios de los productos más estratégicos.

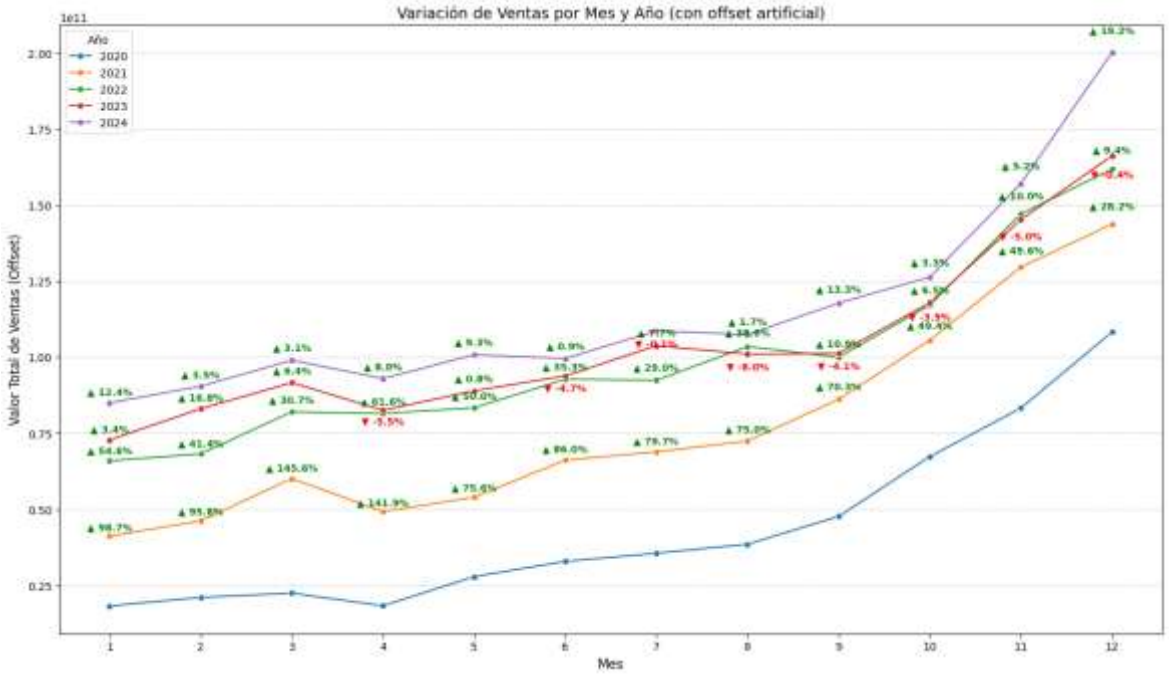

Figura 8.Comportamiento de ventas por mes y año.

Fuente: Elaboración propia.

El gráfico muestra una clara estacionalidad, con incrementos notables en ventas durante los últimos meses de cada año, especialmente en noviembre y diciembre. Además, se observa un crecimiento sostenido de las ventas año tras año, destacando especialmente el incremento en 2024. Las variaciones positivas y negativas interanuales indican que ciertos meses son más sensibles a factores externos como promociones o eventos, lo que debe ser considerado para mejorar la precisión de los modelos de pronóstico.

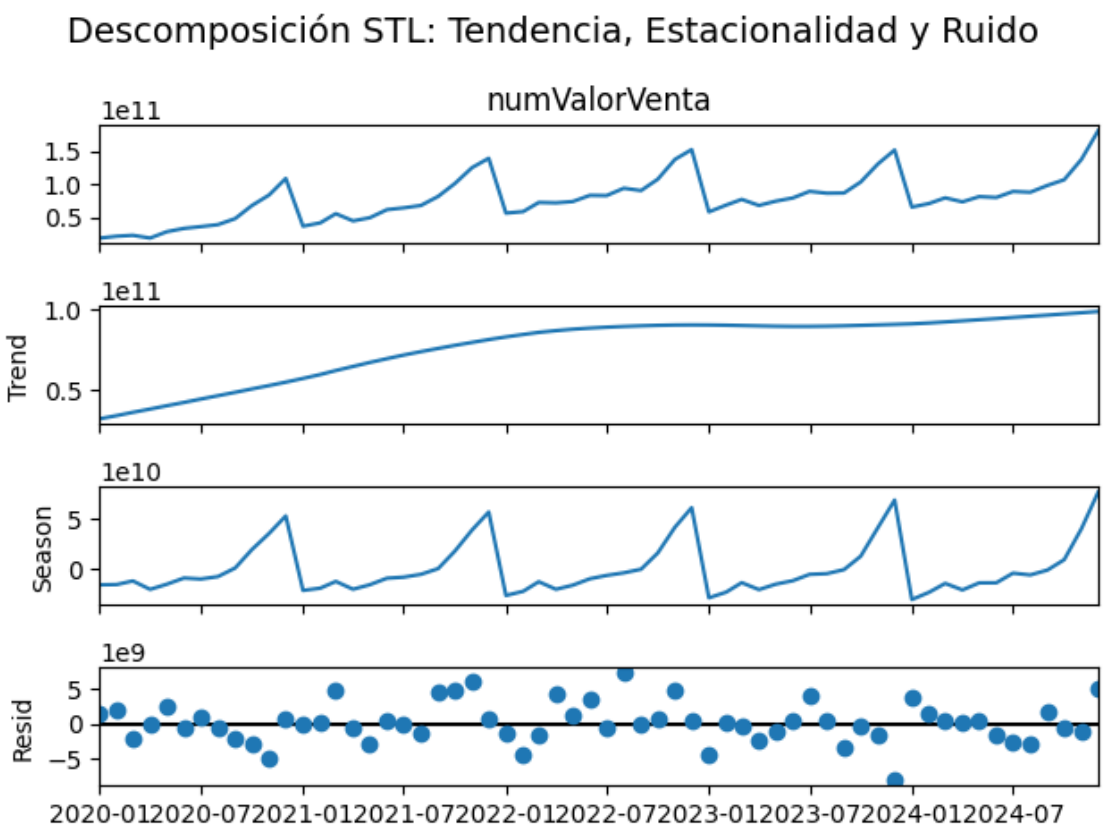

Figura 9.Descomposición, tendencia, estacionalidad y ruido.

Fuente: Elaboración propia.

La descomposición STL revela una tendencia creciente sostenida en las ventas desde 2020 hasta 2024, con estacionalidades claras que se repiten anualmente (picos a fin de año, probablemente diciembre). El componente de ruido muestra fluctuaciones moderadas, sin presencia de grandes anomalías, lo que sugiere que los patrones estacionales y de tendencia son los principales impulsores del comportamiento de las ventas. Esta descomposición es clave para mejorar la precisión de los modelos al capturar adecuadamente las variaciones recurrentes y la tendencia de fondo.

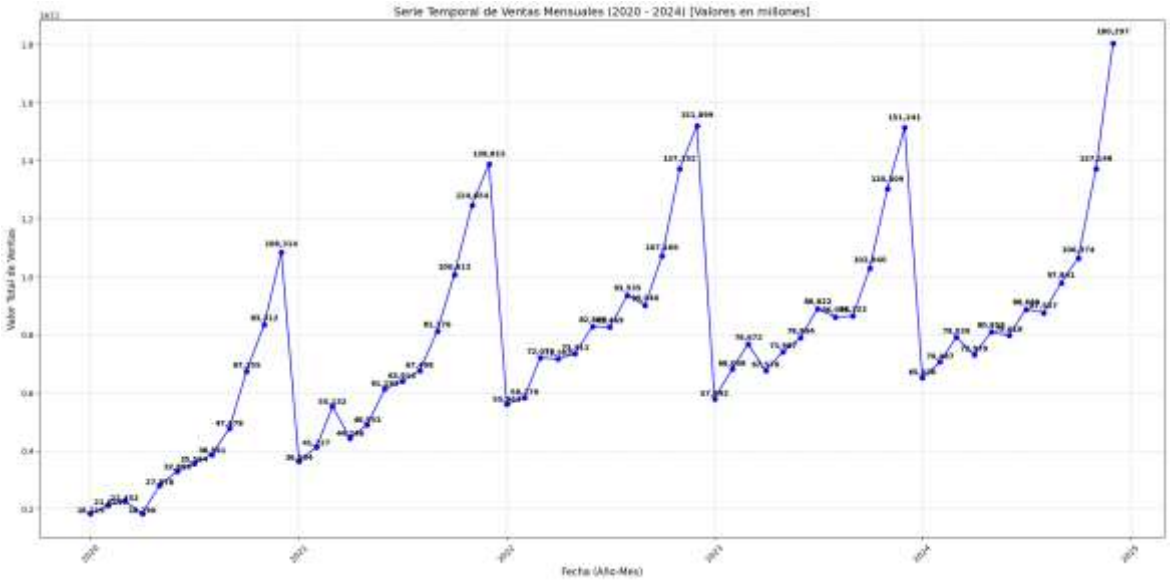

Figura 10.Serie temporal de ventas mensuales.

Fuente: Elaboración propia

La serie temporal muestra un patrón estacional muy claro, con incrementos marcados en las ventas al cierre de cada año (especialmente en diciembre). También se observa una tendencia general de crecimiento sostenido entre 2020 y 2024. Este comportamiento estacional y creciente es característico de industrias como la de bebidas, donde las ventas aumentan en temporadas específicas debido a festividades y promociones.

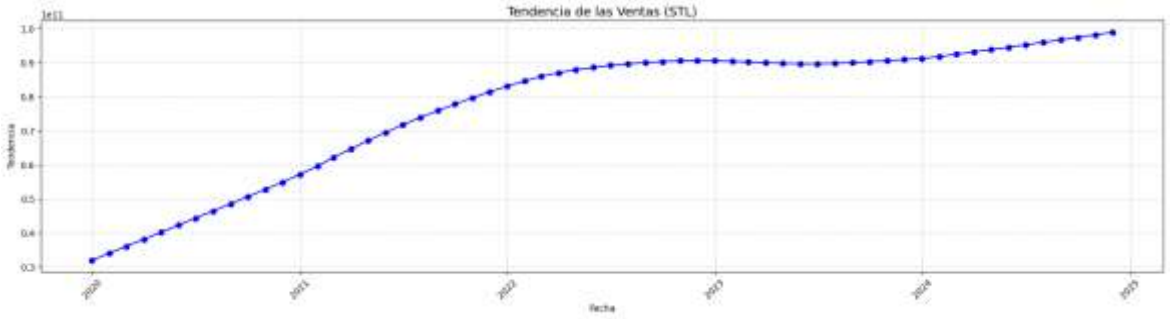

Figura 11.Tendencia de las ventas (STL).

Fuente: Elaboración propia.

El análisis de la tendencia con STL muestra un crecimiento sostenido en las ventas desde 2020, con una ligera estabilización hacia finales de 2022 y comienzos de 2023. Posteriormente, la tendencia retoma un comportamiento ascendente durante 2024. Esta

información es clave para Dislicores, ya que confirma una tendencia positiva a largo plazo, que podría estar asociada al aumento de la demanda general o a la expansión del negocio.

Figura 12.Ruido o residuo de las ventas (STL).

Fuente: Elaboración propia.

El componente de residuo (ruido) presenta fluctuaciones alrededor de la línea base (cero), lo que indica que las desviaciones respecto a la tendencia y estacionalidad son aleatorias y no sistemáticas. Esto sugiere que la mayor parte de la variabilidad en las ventas es explicada por la tendencia y la estacionalidad, mientras que los eventos atípicos son esporádicos. Este comportamiento es positivo, ya que refleja un modelo con buen ajuste y sin sesgos evidentes.

Algunas conclusiones generales al análisis de la tabla de hechos de ventas:

1. Crecimiento sostenido en ventas totales: En este histórico de demanda las ventas reflejan un crecimiento significativo, pasando de aproximadamente $521 mil millones en 2020 a más de $1.1 billones en 2024, lo que refleja una expansión importante del negocio.

2. Estacionalidad marcada y consistente: Se identificó un comportamiento estacional fuerte, con picos de ventas recurrentes en los meses de noviembre y diciembre, coincidiendo con la temporada navideña y promociones especiales.

3. Presencia de valores atípicos justificados: Se detectaron valores negativos en ventas y cantidades facturadas, asociados principalmente a devoluciones y notas crédito. Estos comportamientos son normales en el contexto operativo de la compañía y no representan una anomalía crítica.

4. Concentración de productos clave: El análisis de los productos más vendidos muestra que un grupo reducido de SKUs (top 3) concentra gran parte del volumen de ventas anual, en línea con la regla de Pareto (80/20), lo que destaca la importancia estratégica de estas referencias para la operación.

Tendencia positiva y estabilidad del modelo: La descomposición STL evidenció una tendencia de ventas positiva y estable a lo largo de los años, mientras que el componente de residuo mostró fluctuaciones aleatorias sin patrones sistemáticos, lo que confirma la calidad y robustez de los datos para la construcción de modelos predictivos.

6. Implementación y comparación de modelos

La implementación y comparación de modelos es una etapa esencial en proyectos de pronóstico, ya que permite construir, ajustar y validar diversos algoritmos para resolver un problema específico, como la predicción de la demanda. Esta fase no solo implica entrenar modelos, sino también optimizar sus parámetros y automatizar su ejecución (Kuhn & Johnson, 2013).

Posteriormente, se comparan los resultados mediante métricas objetivas como MAE, SESGO% y SCORE%, lo que garantiza la selección del modelo más preciso y robusto para las necesidades del negocio (Hyndman & Athanasopoulos, 2018). Esta evaluación asegura una toma de decisiones más confiable y basada en datos.

6.1. Modelos de series de tiempo

En esta sección se presentan los resultados de la implementación de varios modelos de series de tiempo aplicados al pronóstico de la demanda de Dislicores. Estos modelos fueron seleccionados por su capacidad para capturar patrones de comportamiento en datos históricos, tales como tendencia, estacionalidad y variaciones aleatorias. La elección de los

modelos responde a la necesidad de comparar enfoques tradicionales y avanzados en la predicción de la demanda.

6.1.1. Modelo de línea base actual: Proceso Supply Chain como Benchmark

En este apartado se documenta la réplica del modelo de pronóstico desarrollado por el director de Supply Chain de Dislicores, el cual constituye el Benchmark oficial contra el que se evaluarán los modelos avanzados propuestos en este proyecto. Este modelo, construido y aplicado en hojas de cálculo Excel, ha sido la base de las decisiones estratégicas de la compañía en materia de planeación de demanda y gestión de inventarios, entregando resultados aceptables para el negocio.

El modelo es un enfoque híbrido que combina técnicas estadísticas y reglas de negocio para suavizar la demanda y realizar la predicción de productos claves del portafolio (segmento Pareto). El proceso incluye:

1. Uso de dos años históricos como base de datos.
2. Cálculo para cada SKU, de forma anual, de métricas como: promedio, desviación estándar, mínimo, máximo y coeficiente de variación ($CV = \sigma/\mu$).
3. Clasificación de cada referencia en categorías X, Y o Z en función de su CV:
    - X: $CV \leq 1$
    - Y: $1 < CV \leq 2$
    - Z: $CV > 3$
4. Suavización de datos para categorías Y y Z reemplazando el valor máximo por el promedio.
5. Combinación de ambos años históricos promediando cada mes.
6. Cálculo de índices estacionales con base en el promedio general de cada referencia.
7. Desestacionalización de la serie suavizada utilizando dichos índices.
8. Se realiza una proyección paralela de dos modelos:
    - Una regresión lineal simple aplicada a la serie desestacionalizada.
    - Un modelo Holt-Winters aplicado sobre la serie suavizada.

Selección del modelo:

- Para cada mes de la proyección, se comparan las predicciones de la regresión lineal y Holt-Winters.
- El modelo con menor error (menor distancia respecto a la serie real) es seleccionado como el valor final del pronóstico para dicho mes.

Modelos derivados del proceso de replicación de Supply Chain: De este enfoque se obtienen tres conjuntos de resultados o "modelos" para efectos de la comparación posterior:

- Modelo Supply Chain (Benchmark): Selección del mejor valor mensual entre la regresión lineal y Holt-Winters.
- Regresión Lineal Supply Chain: Predicciones resultantes exclusivamente de la regresión lineal aplicada.
- Holt-Winters Supply Chain: Predicciones resultantes únicamente del modelo Holt-Winters.

Este enfoque ha sido de gran utilidad para Dislicores, pero el presente proyecto busca validar si la incorporación de modelos avanzados, como XGBoost, LightGBM y MSTL con regresión polinómica, puede superar en precisión a este proceso tradicional.

6.1.2. Supply Chain

A continuación, se presentan los resultados de la evaluación del modelo de pronóstico actual utilizado por el área de Supply Chain, que sirve como línea base para la comparación con los modelos propuestos en este estudio.

Resultados del modelo Supply Chain:

- MAE% global: 29,37% El modelo presenta un nivel de error absoluto medio del 29,37%. Este valor es relativamente elevado, lo que indica que, en promedio, las predicciones realizadas presentan una diferencia importante respecto a la demanda real. Aunque ha sido funcional para la organización, este porcentaje sugiere que existe una oportunidad clara de mejorar la precisión del pronóstico mediante modelos alternativos.

- Sesgo% global: -6,01% El modelo tiende a subestimar la demanda real en un 6,01%, evidenciando una desviación sistemática hacia valores por debajo de la realidad. Aunque este valor está cercano al rango recomendado de ±5% para evitar sesgos significativos, es necesario tomar precauciones, ya que la subestimación puede afectar negativamente la disponibilidad de inventario y la capacidad de respuesta ante la demanda.
- Score global: 35,38% El Score global, que integra el MAE% y la magnitud del sesgo absoluto, alcanza un 35,38%. Esta métrica refleja que el modelo posee un margen de error considerable y una tendencia subestimada en sus proyecciones.

Conclusión preliminar: El modelo de Supply Chain es efectivo como punto de partida, pero sus resultados dejan espacio para explorar modelos más sofisticados que puedan reducir tanto el error como el sesgo, optimizando la planificación de la demanda y la eficiencia operativa de la organización.

6.1.3. Regresión línea Supply Chain

El modelo de Regresión Lineal, empleado de forma individual dentro de la metodología de Supply Chain, ofrece las siguientes métricas de desempeño:

Resultados del modelo de Regresión Lineal Supply Chain:

- MAE% global: 27,9% El error absoluto medio de este modelo es del 27,9%, lo que indica una desviación promedio cercana al 28% respecto a la demanda real. Aunque no es una precisión óptima, representa una ligera mejora respecto al modelo combinado Supply Chain, manteniéndose por debajo del umbral del 30%.
- Sesgo% global: 1,87% El sesgo positivo muestra que el modelo tiende levemente a sobreestimar la demanda, aunque el valor de 1,87% se encuentra dentro de los límites aceptables (±5%), lo que sugiere que no presenta un sesgo sistemático relevante.
- Score global: 29,77% El score global de 29,77% refleja un nivel de error moderado con una ligera tendencia a sobreestimación. Si bien es mejor que el resultado obtenido por el modelo combinado Supply Chain, aún se observa margen de mejora.

Conclusión preliminar: El modelo de Regresión Lineal presenta un comportamiento más balanceado en términos de sesgo y mejora ligeramente el MAE% respecto al modelo base de Supply Chain. Sin embargo, sigue sin alcanzar niveles óptimos de precisión, por lo que sugiere la necesidad de explorar enfoques más sofisticados que logren una reducción adicional del error.

6.1.4. Holt Winters Supply Chain

El modelo Holt-Winters, utilizado como alternativa dentro de la metodología de Supply Chain, muestra los siguientes resultados en cuanto a precisión y sesgo:

Resultados del modelo Holt-Winters Supply Chain:

- MAE% global: 33,94% El error absoluto medio se sitúa en 33,94%, reflejando un nivel de error relativamente alto en las predicciones. Esto sugiere que, en promedio, el modelo no logra capturar con alta precisión la demanda real. Un MAE% por debajo del 30% sería más deseable para este tipo de aplicación.

- Sesgo% global: 5,82% El sesgo positivo de 5,82% muestra que el modelo tiende ligeramente a sobrestimar la demanda. Sin embargo, este valor supera el umbral recomendado de ±5%, lo que podría afectar la eficiencia de la gestión de inventarios al proyectar demandas ligeramente infladas.
- Score global: 39,76% Este score refleja la combinación de un error absoluto moderadamente alto y un sesgo que supera el límite sugerido. El resultado indica que el modelo es menos eficiente que la Regresión Lineal, lo que reduce su competitividad frente a otros enfoques más avanzados.

Conclusión preliminar: El modelo Holt-Winters, aunque útil para capturar patrones de estacionalidad y tendencia, presenta una mayor desviación y sesgo que los modelos anteriores. Esto refuerza la necesidad de implementar y evaluar modelos más sofisticados que integren variables exógenas y ajusten mejor las fluctuaciones de la demanda.

6.1.5. Promedio Móvil Simple

El modelo de Promedio Móvil Simple (SMA) fue implementado como referencia para aquellas referencias que no cuentan con un mínimo de dos años de histórico, limitando la validación de estacionalidad.

Resultados del modelo SMA:

- MAE% global: 37,52% El modelo presenta un error absoluto medio elevado, lo que sugiere que, en promedio, las predicciones difieren significativamente de la demanda real. Esta alta desviación lo posiciona como un modelo básico y menos preciso para la proyección de demanda.
- Sesgo% global: -5,47% El sesgo negativo indica una tendencia a subestimar la demanda en un 5,47% promedio. Si bien se encuentra cercano al umbral recomendado de ±5%, esta subestimación puede impactar negativamente en la planificación de inventarios y la atención a la demanda real.
- Score global: 42,99% El score, resultado de la combinación del MAE% y el sesgo, es relativamente alto, lo que refuerza la conclusión de que el modelo no logra un ajuste adecuado y presenta limitaciones como predictor confiable para la toma de decisiones operativas.

Nota: Este modelo se mantiene en la cartera de opciones como alternativa para referencias con poco histórico disponible, donde modelos estacionales o complejos no pueden ser implementados debido a la falta de datos suficientes.

Conclusión preliminar: El modelo de Promedio Móvil Simple (SMA), aunque es una herramienta útil en escenarios donde el volumen histórico de datos es limitado, demuestra un desempeño inferior en comparación con otros modelos más robustos. Su MAE% elevado (37,52%) y su score global alto (42,99%) reflejan su baja capacidad para capturar patrones complejos como estacionalidades o tendencias. No obstante, es una opción válida en contextos donde la carencia de datos históricos impide aplicar modelos más sofisticados. Esta situación resalta la importancia de contar con series de tiempo más extensas para poder implementar modelos con mejor capacidad predictiva.

6.1.6. MSTL con Regresión Polinómica

- MAE% global: 27,78% El MAE% global de 27,78% indica que el modelo tiene un error moderado en sus predicciones. Si bien no es el más bajo posible, se encuentra por debajo del umbral crítico del 30%, lo que sugiere una precisión aceptable dentro del contexto de la demanda en Dislicores.

- Sesgo% global: -0,18% El sesgo de -0,18% muestra que el modelo es prácticamente neutro, con una leve tendencia a subestimar la demanda. Este valor es muy cercano a cero, lo que es ideal, ya que reduce el riesgo de errores sistemáticos que puedan sesgar la toma de decisiones.

- Score global: 27,96% Este score global es coherente con las métricas anteriores y refleja una combinación entre la precisión y la ausencia de sesgo. Si bien sigue existiendo un margen de error moderado, el modelo muestra un buen equilibrio entre error absoluto y tendencia.

Conclusión preliminar: El modelo MSTL con regresión polinómica ha logrado resultados estables y balanceados, con un sesgo prácticamente nulo y un MAE% aceptable. Esto lo posiciona como una alternativa competitiva frente a los modelos tradicionales y a las técnicas usadas actualmente por Supply Chain. Además, su capacidad para capturar patrones de estacionalidad y tendencia lo convierte en un modelo útil para contextos donde la demanda presenta variaciones estacionales importantes.

6.1.7. Holt-Winters

Este modelo se ejecutó sin aplicar los preprocesamientos adicionales que fueron implementados en los modelos desarrollados por la dirección de Supply Chain (como suavizados previos o desestacionalización).

- MAE% global: 31,03% El MAE% global de 31,03% sugiere que el modelo presenta un error moderado en sus predicciones. Aunque no es el peor desempeño observado, este nivel de error indica que el modelo podría beneficiarse de técnicas

adicionales de ajuste o de una mayor integración de variables externas.

- Sesgo% global: -0,83% El sesgo negativo es leve, con solo un -0,83%, lo que indica que el modelo tiende a subestimar ligeramente la demanda, pero sin ser una desviación sistemática grave. Este sesgo está dentro de rangos aceptables, mostrando que el modelo está relativamente bien calibrado en cuanto a dirección de error.
- Score global: 31,85% El score global, que combina tanto el MAE% como el sesgo, muestra un rendimiento general que podría ser mejorado. Aunque el sesgo es bajo, el MAE elevado impacta negativamente el score, señalando que la precisión general aún es limitada.

Conclusión preliminar: El modelo Holt-Winters presenta una ligera subestimación y un error moderado. Si bien es una herramienta clásica y confiable para series con estacionalidad, en este caso, al no haber sido complementado con preprocesamientos adicionales, presenta un rendimiento intermedio en comparación con otros modelos evaluados. Se recomienda considerar su integración como base para modelos híbridos o complementarla con enfoques más avanzados en contextos de alta variabilidad.

6.2. Modelos de machine learning

En esta sección se abordan modelos de Machine Learning aplicados al pronóstico de la demanda, específicamente técnicas basadas en algoritmos de Gradient Boosting. A diferencia de los modelos tradicionales de series temporales, los enfoques de Machine Learning permiten incorporar variables exógenas y manejar relaciones no lineales, mejorando la capacidad predictiva en escenarios complejos.

Se destacan dos algoritmos ampliamente utilizados por su eficiencia y precisión en tareas de predicción:

- XGBoost: Un algoritmo de boosting que se ha consolidado como una herramienta potente para la predicción de demanda gracias a su capacidad para minimizar el error y manejar grandes volúmenes de datos con alta dimensionalidad.

- LightGBM: Una alternativa optimizada que mejora la velocidad de entrenamiento y reduce el uso de memoria, siendo especialmente útil para bases de datos voluminosas y entornos productivos donde la eficiencia es clave.

Ambos modelos han sido aplicados sobre la misma serie temporal que los modelos tradicionales, pero con la ventaja de integrar variables estacionales y exógenas como promociones, clima y festividades, logrando capturar de manera más robusta las fluctuaciones de la demanda.

6.2.1. Definición de búsqueda de Hiperparámetro

En este proceso, se construyen 10 combinaciones de hiperparámetro con el objetivo de optimizar el modelo. Para ello, se utiliza el método bayesiano, una técnica que permite encontrar las mejores combinaciones de hiperparámetro de manera eficiente, evitando la evaluación de todas las opciones posibles, lo que reduciría significativamente el costo computacional.

A diferencia de una búsqueda aleatoria o exhaustiva, la búsqueda bayesiana se basa en la exploración informada del espacio de hiperparámetro, utilizando los resultados de iteraciones previas para enfocar la búsqueda en las regiones más prometedoras (Snoek, Larochelle & Adams, 2012). Esto permite priorizar aquellas configuraciones que presentan mejor desempeño, reduciendo el tiempo de entrenamiento y maximizando la eficiencia.

Gracias a esta metodología, el ajuste de hiperparámetro se vuelve más rápido y efectivo, logrando una optimización robusta del modelo sin necesidad de agotar recursos computacionales en combinaciones poco relevantes.

6.2.2. Generación de atributos

A diferencia de los modelos tradicionales de series de tiempo, los modelos de machine learning no poseen una comprensión inherente de la secuencialidad de los datos. Para que estos modelos puedan capturar la estacionalidad y patrones temporales, es fundamental

generar atributos diseñados específicamente para reflejar estas características (Hyndman & Athanasopoulos, 2018).

Entre los atributos más relevantes se encuentran:

- **Lags temporales** (lag_1m, lag_12m, etc.): Representan la demanda de períodos anteriores y ayudan al modelo a identificar patrones repetitivos.
- **Promedios estacionales**: Cálculo del promedio de la demanda en períodos específicos de los últimos años, lo que permite al modelo reconocer patrones cíclicos.
- **Mediana estacional**: Similar al promedio, pero más robusta ante valores atípicos.
- **Máximos y mínimos históricos**: Determinan los valores extremos en los últimos tres años, ayudando al modelo a identificar rangos de comportamiento de la demanda.
- **Desviación estándar estacional**: Mide la variabilidad de la demanda en períodos pasados, útil para detectar fluctuaciones y estabilidad.
- **Ciclos estacionales**: Identifican si la demanda sigue patrones específicos dentro de cada período (meses, trimestres, etc.).

Además, se integran **variables exógenas**, como:

- El **número del mes** (captura de estacionalidad mensual).
- **Indicadores de días festivos y eventos especiales**, que afectan directamente la demanda.

Según Han, Pei y Kamber (2011), la ingeniería de atributos adecuada permite que los modelos de machine learning capturen mejor las relaciones temporales y mejoren su capacidad predictiva en escenarios complejos como la planificación de la demanda. Incorporar estas variables asegura pronósticos más robustos y ajustados a la realidad del negocio.

6.2.3. XGBoost local

XGBoost es un algoritmo de machine learning basado en árboles de decisión que mejora el rendimiento de los modelos predictivos mediante la técnica de boosting. En este enfoque, los árboles se construyen de manera secuencial, donde cada nuevo árbol corrige los errores

cometidos por los anteriores, incrementando así la precisión del modelo (Chen & Guestrin, 2016).

- Ventajas de XGBoost:
- Procesa grandes volúmenes de datos de manera eficiente.
- Integra mecanismos de regularización que reducen el riesgo de sobreajuste.
- Aprovecha la computación paralela, acelerando su ejecución.
- Es especialmente eficaz para tareas de clasificación, regresión y predicción de series temporales, incluso en entornos con alta dimensionalidad y ruido (Brownlee, 2016).

Aplicación Local: En este proyecto, XGBoost se implementa de manera individual para cada SKU (modelo local), lo que implica:

- Un ajuste de hiperparámetro específico para cada SKU.
- Generación de pronósticos personalizados según los datos históricos de cada producto.
- Evaluación individualizada de los resultados para determinar la precisión SKU por SKU.

Si bien este enfoque garantiza una alta personalización y precisión, el costo computacional es elevado, ya que la ejecución completa puede tardar entre 4 y 5 horas debido al volumen de datos y la cantidad de SKUs.

Resultados:

- MAE% global: 11,66% El modelo presenta un error bajo en sus predicciones, lo que refleja una alta precisión al capturar la demanda real de los productos.
- Sesgo% global: -3,08% El modelo muestra una leve tendencia a subestimar la demanda, pero se mantiene dentro de un rango aceptable (±5%), indicando que las predicciones son relativamente balanceadas
- Score global: 14,75% El score refleja un excelente equilibrio entre precisión y sesgo, posicionando al XGBoost local como una opción sólida para la demanda a nivel de SKU

Conclusión preliminar: El XGBoost Local sigue siendo uno de los modelos más destacados del análisis, con un MAE% de 11,66% y un sesgo controlado de -3,08%, lo que demuestra

su capacidad para ofrecer predicciones precisas y balanceadas a nivel SKU. A pesar de su mayor demanda computacional, es altamente recomendable en contextos donde es crucial personalizar los pronósticos para cada producto, especialmente en negocios con alta diversidad de SKUs.

6.2.4. XGBoost global

A diferencia del enfoque XGBoost Local, que entrena un modelo específico para cada SKU, el XGBoost Global genera un único modelo que considera todos los SKUs de forma conjunta. Esto reduce significativamente el tiempo de procesamiento, pasando de 4-5 horas (modelo local) a 45 minutos - 1 hora, lo que lo convierte en una alternativa eficiente para entornos donde la rapidez es clave.

Ventaja principal: El modelo global permite explotar la capacidad de generalización de XGBoost, capturando patrones comunes entre las referencias de productos.

Limitación: Al tratar todas las referencias de manera agregada, se pierde parte de la personalización que ofrecen los modelos individuales. Esto puede impactar en la capacidad del modelo para captar particularidades de productos específicos, afectando la precisión a nivel SKU.

¿Cuándo es útil? El XGBoost Global es ideal en situaciones donde:

- Se necesita una respuesta rápida en entornos productivos.
- La variabilidad entre SKUs es baja o moderada.
- Se prioriza eficiencia computacional sobre personalización completa.

Resultados:

- MAE% global: 10,85% Un resultado excelente, logrando la menor desviación media entre las predicciones y los valores reales entre modelos probados
- Sesgo% global: -2,45% Este sesgo es ligeramente negativo, lo que sugiere que el modelo tiende a subestimar la demanda de manera moderadas, pero dentro de los límites aceptables.
- Score global: 13,31% El balance entre el bajo MAE% y el sesgo controlado posiciona al XGBoost Global como uno de los modelos más competitivos de todo el estudio.

Conclusión preliminar: El XGBoost Global presenta el mejor desempeño general del estudio, con un MAE% de 10,85% y un sesgo muy bajo de -2,45%, lo que lo convierte en una excelente opción para escenarios donde se prioriza la eficiencia y rapidez de ejecución sin comprometer la calidad de las predicciones. Sin embargo, en productos con patrones muy particulares o atípicos, el enfoque Local podría seguir siendo superior.

6.2.5. LigthGBM local

LightGBM Local es una variante del algoritmo de boosting que, a diferencia de la versión global, entrena un modelo individual para cada SKU de forma aislada. Este enfoque busca capturar mejor los patrones particulares de cada producto, ajustando los hiperparámetro para cada serie de manera personalizada.

Sin embargo, al aplicar este enfoque, el tiempo de ejecución aumenta en comparación con el modelo global, ya que se repite el proceso de entrenamiento para cada referencia de forma independiente. Aunque la expectativa de un modelo local es mejorar la precisión al adaptarse a las características específicas de cada SKU, en este caso particular, el desempeño fue inferior al de otros enfoques evaluados.

Características de la implementación:

- Entrenamiento individual de un modelo LightGBM por cada SKU.
- Ajuste de hiperparámetro específico para cada serie.
- Tiempo de ejecución aproximado: 1,5 a 2 horas, dependiendo del volumen de SKUs y la complejidad de cada serie.

Resultados:

- MAE% global: 20,75% El modelo presenta un nivel de error más alto, lo que indica una menor precisión en sus predicciones.
- Sesgo% global: -18,25% Este sesgo elevado refleja una fuerte tendencia a subestimar la demanda, lo cual puede afectar la planificación, provocando posibles faltantes de inventario.
- Score global: 39,00% Un score relativamente elevado que refleja la combinación de un error absoluto significativo y un sesgo importante. El alto score sugiere que este

modelo no es la mejor opción para el contexto evaluado, ya que combina un error y un sesgo considerables.

Conclusiones preliminares:

- Aunque LightGBM es reconocido por su eficiencia, en este caso la versión Local mostró limitaciones importantes, con un MAE% de 20,75% y un sesgo de -18,25%, reflejando una subestimación considerable de la demanda.
- Esta fuerte subestimación podría afectar la gestión operativa, generando desabastecimientos o problemas en la planificación logística.
- A pesar de ser un modelo rápido y eficiente en otras aplicaciones, LightGBM Local, en este contexto, requiere ajustes adicionales o podría ser reemplazado por enfoques más robustos como XGBoost Local.

6.2.6. LigthGBM global

LightGBM es un algoritmo de machine learning que aplica la técnica de boosting mediante árboles de decisión, optimizado para ofrecer mayor rapidez y eficiencia en comparación con otros métodos como XGBoost. Su principal ventaja radica en su capacidad para dividir los datos de forma más eficiente (método "leaf-wise") y reducir el uso de memoria, lo que lo hace especialmente útil para grandes volúmenes de datos.

En este caso, LightGBM se implementa en un modelo global, es decir, se entrena un único modelo para todas las referencias (SKUs) combinadas. Esto permite aprovechar la velocidad de entrenamiento del algoritmo y obtener resultados más rápidos en comparación con otros modelos individualizados.

Características de la implementación:

- Ajuste global de hiperparámetro para todas las referencias.
- El modelo procesa todos los SKUs simultáneamente, reduciendo significativamente los tiempos de ejecución.
- La ejecución completa oscila entre 1 hora y 90 minutos, dependiendo del tamaño y la complejidad del dataset.

Observaciones:

- Aunque LightGBM es más rápido que XGBoost Local, la precisión obtenida no es tan alta como en modelos más personalizados.
- Este modelo es una opción atractiva cuando la velocidad y eficiencia computacional son prioridades, aunque puede requerir ajustes adicionales para mejorar la precisión en contextos específicos.

Resultados:

MAE% global: 27,51% Un error absoluto moderado, indicando que el modelo tiene un margen de mejora respecto a otros enfoques más precisos.

Sesgo% global: 3,05% El modelo presenta una leve tendencia a sobreestimar la demanda, aunque se mantiene dentro de un rango aceptable (±5%).

Score global: 30,56% El score global sugiere un rendimiento razonable pero mejorable, especialmente si se prioriza la precisión por encima de la eficiencia computacional.

Conclusión preliminar: LightGBM Global ofrece un balance adecuado entre rapidez y rendimiento, siendo útil para escenarios donde se requiere una solución rápida sin comprometer en exceso la precisión. Sin embargo, para demandas más personalizadas o productos con comportamiento muy específico, otros enfoques como XGBoost Local podrían ser más adecuados.

6.3. Comparación de resultados de modelos

La comparación de modelos es una etapa clave para determinar cuál enfoque de pronóstico ofrece el mejor desempeño en función de métricas objetivas. En este caso, se evaluaron diversos modelos de series de tiempo y machine learning aplicados a la predicción de la demanda de Dislicores. La evaluación se realizó utilizando métricas como MAE%, Sesgo% y Score Global para cada SKU y de manera agregada. A continuación, se presentan los resultados obtenidos, destacando el rendimiento de cada modelo, su impacto a nivel global y la distribución del mejor modelo por referencia.

6.3.1. Evaluación global del forecast

La aplicación del modelo de Forecast en su conjunto global, es decir, seleccionando el mejor modelo para cada SKU, arroja las siguientes métricas globales:

- MAE% global: 20,84%
- Sesgo global: 0,12%
- Score global: 20.96%

Este resultado global representa una mejora sustancial respecto al modelo base actualmente empleado por el equipo de Supply Chain (SC). Mientras que el modelo SC arrojaba un Score Global de 35,38% y un MAE% de 29,37%, la implementación del modelo combinado de forecast permitió reducir el MAE% a 20,84% y el Score Global a 20,96%. Esta diferencia refleja una optimización considerable en la capacidad predictiva, disminuyendo los errores absolutos y mejorando el balance general entre error y sesgo.

La reducción de más de 14 puntos porcentuales en el Score Global implica una mayor precisión y confiabilidad en la planificación de la demanda, lo que se traduce en mejores decisiones operativas y estratégicas para la organización. Esta mejora refuerza la eficacia del enfoque basado en machine learning y técnicas avanzadas de modelado frente a métodos tradicionales.

6.3.2. Mejor modelo

El modelo XGBoost Local se posiciona como la opción más robusta y precisa en el contexto del pronóstico de la demanda de Dislicores. Su capacidad para generar modelos individualizados para cada SKU permite capturar de manera más efectiva las particularidades y patrones únicos de cada referencia. Esto es especialmente valioso en industrias con alta diversidad de productos, como el sector de bebidas, donde la estacionalidad y la volatilidad de la demanda pueden variar significativamente entre SKUs.

Resultados destacados:

- MAE% Global: 11,66% El error absoluto medio es el más bajo entre todos los modelos evaluados. Esto significa que, en promedio, las predicciones del modelo

difieren solo un 11,66% de la demanda real, lo que se traduce en una alta precisión y menor riesgo de desviaciones significativas en la planificación de inventarios.

- Sesgo% Global: -3,08% El sesgo negativo indica una ligera tendencia a subestimar la demanda, aunque se mantiene dentro de un rango aceptable (<±5%). Esta característica es relevante ya que ayuda a evitar una sobreestimación de inventarios, contribuyendo a una gestión más eficiente de los recursos logísticos y de almacenamiento.
- Score Global: 14,75% Este valor combinado (que integra MAE% y Sesgo%) confirma el sólido desempeño del modelo, consolidándolo como el más equilibrado entre precisión y estabilidad. En comparación con otros modelos, XGBoost Local es el que logra minimizar tanto el error absoluto como la desviación sistemática.

Ventajas clave del XGBoost Local:

- Personalización por SKU: Al construir un modelo específico para cada referencia, el XGBoost Local es capaz de capturar las dinámicas individuales de cada producto, desde los más estables hasta los de comportamiento más errático.
- Alta precisión: Gracias al enfoque granular y la optimización de hiperparámetro por SKU, logra superar ampliamente a los modelos globales y a los métodos tradicionales.
- Desempeño superior en el 56% de los SKUs: Según la distribución de resultados por SKU, XGBoost Local es el mejor modelo en más de la mitad de las referencias, consolidando su rol como la solución preferente.

El XGBoost Local no solo lidera en métricas de error y sesgo, sino que también aporta un valor estratégico a la operación, permitiendo una planificación más precisa y con menores riesgos asociados al desabastecimiento o al sobre stock.

6.3.3. Distribución de mejor modelo por SKU

La siguiente gráfica muestra cómo se distribuyen los mejores modelos seleccionados por SKU:

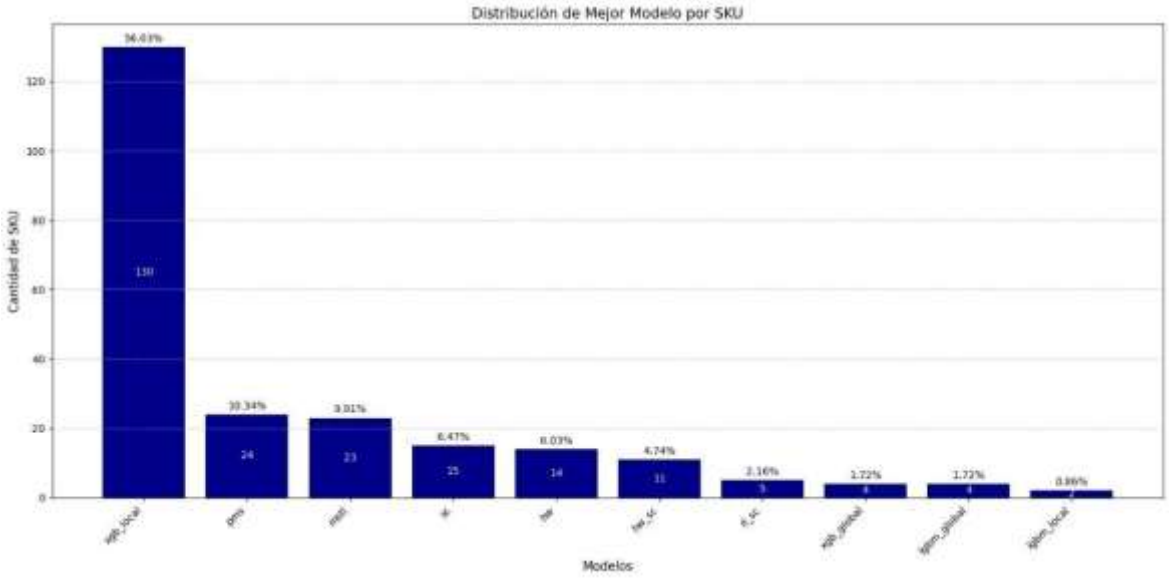

Figura 13.Distribución de mejor modelo por SKU.

Fuente: Elaboración propia

- XGBoost Local se posiciona como el modelo más utilizado, seleccionado en el 56,03% de los SKUs analizados.
- Otros modelos como Promedio Móvil Simple y MSTL ocupan posiciones secundarias con un peso menor en la distribución.

6.3.4. Modelo con peor desempeño

El modelo que presenta el rendimiento más bajo en términos de precisión y balance es el LightGBM Local:

- **MAE% Global:** 20,75%
- **Sesgo% Global:** -18,25%
- **Score Global:** 39,00%

A pesar de ser un modelo rápido y eficiente en otros contextos, LightGBM Local mostró una tendencia marcada a subestimar la demanda (sesgo de -18,25%), comprometiendo la confiabilidad de las predicciones y afectando la toma de decisiones operativas.

6.3.5. Versus: Modelos SC (Supply Chain) vs Modelos optimizados

La comparación de resultados es una etapa crucial para evaluar la efectividad de los modelos propuestos frente al modelo tradicional utilizado por el equipo de Supply Chain (SC). En este análisis se busca determinar cómo los modelos optimizados mejoran la precisión de las predicciones y la cobertura de SKU, en comparación con el modelo SC, que ha sido la referencia base en la organización.

Para evaluar la mejora, se aplican dos métricas clave:

1. Variación en la cantidad de SKU impactadas:
    a. Se calcula cómo cada modelo optimizado afecta la cobertura de SKU en comparación con la cantidad de SKU gestionadas bajo el modelo SC.
    b. Formula:

$$Variación\ SKU\ (\%) = \frac{(Cant.SKU\ Modelo - Cant.SKU\ SC)}{Cant.SKU\ SC} \times 100$$

    c. Esta métrica muestra si el nuevo modelo logra abarcar más SKU que el modelo base o si, por el contrario, reduce la cantidad de SKU impactadas.

2. Variación en el score global:
    a. El Score Global es una métrica compuesta que combina el MAE% y el Sesgo% de cada modelo, reflejando su desempeño general. Se compara el Score Global de cada modelo frente al Score Global del modelo SC.
    b. Formula:

$$Variación\ Score\ (\%) = \frac{(Score\ Modelo - Score\ SC)}{Score\ SC} \times 100$$

    c. Una variación negativa es positiva, ya que indica una reducción en el error total.
3. ¿Por qué es relevante esta comparación?
    a. Evaluar cobertura: Se identifica si los modelos optimizados logran cubrir más referencias o SKU clave.
    b. Medir la eficiencia predictiva: Se determina qué modelos no solo abarcan más SKU, sino también generan mejores resultados en términos de precisión y balance entre error y sesgo.
    c. Tomar decisiones estratégicas: Con esta información se facilita la selección del mejor modelo según el contexto operativo de Dislicores.

Tabla 3.Comparación de modelos optimizados vs Modelo SC.

Modelo	Cant. SKU (Modelo)	Cant. SKU SC	Variación SKU (%)	Score Global (Modelo)	Score Global SC	Variación Score (%)
XGBoost Local	130	15	+766.7%	14.75%	35.38%	-58.3%
MSTL + Regresión Polinómica	23	15	+53.3%	27.96%	35.38%	-20.9%
Holt-Winters (sin	14	15	-6.7%	31.85%	35.38%	-10.0%

preprocesar.)						
Holt-Winters SC	11	15	-26.7%	39.76%	35.38%	+12.4%
LightGBM Global	4	15	-73.3%	30.56%	35.38%	-13.6%
LightGBM Local	2	15	-86.7%	39.00%	35.38%	+10.2%
XGBoost Global	4	15	-73.3%	13.31%	35.38%	-62.4%
Promedio Móvil Simple	24	15	+60.0%	42.99%	35.38%	+21.5%
Regresión Lineal SC	5	15	-66.7%	29.77%	35.38%	-15.8%

Conclusiones preliminares:

- XGBoost Local y XGBoost Global muestran la mayor reducción en el Score Global, logrando entre un 58% y 62% de mejora respecto al modelo SC. Además, XGBoost Local cubre 130 SKU, un aumento de +766.7% respecto al SC.
- Promedio Móvil Simple y LightGBM Local presentan mayor Score que el SC, lo que sugiere un peor desempeño predictivo a pesar de que Promedio Móvil Simple cubre más SKU.
- La mayoría de los modelos optimizados, especialmente los basados en XGBoost, no solo mejoran la precisión del pronóstico, sino que también logran impactar más SKU estratégicas para la compañía.

La comparación detallada de los modelos revela una mejora significativa al implementar enfoques optimizados frente al modelo tradicional de Supply Chain (SC). En particular, los

modelos basados en XGBoost (tanto Local como Global) lograron reducciones sustanciales en el Score Global, superando ampliamente al modelo SC en términos de precisión y balance entre error y sesgo.

Además, el XGBoost Local no solo es el modelo más preciso, sino que también amplió la cobertura de SKU de forma notable, impactando a una mayor cantidad de referencias críticas para la operación. Por otro lado, modelos como el Promedio Móvil Simple o LightGBM Local, aunque útiles en contextos específicos, mostraron limitaciones frente a los modelos avanzados en términos de desempeño global.

Estos hallazgos demuestran que la aplicación de modelos avanzados de machine learning, junto con una adecuada generación de atributos y ajuste de hiperparámetro, permite optimizar la gestión de la demanda, mejorando la toma de decisiones y fortaleciendo la competitividad operativa de Dislicores.

7. Implementación y selección de modelos

Una vez desarrollados y comparados los distintos modelos de pronóstico, es esencial implementar y evaluar su desempeño en un entorno real. Esta etapa permite verificar no solo la precisión teórica obtenida durante la experimentación, sino también su aplicabilidad y robustez frente a datos reales de negocio.

El sistema de forecasting implementado para Dislicores ha sido ejecutado de manera automatizada durante el primer mes del año 2025, posterior a esto, seguirá una lógica de ventanas móviles. Esto significa que, cada mes, se reentrena el modelo utilizando los datos históricos más recientes y se generan pronósticos actualizados para los siguientes 12 meses.

Con este enfoque, no solo aseguramos que el modelo esté siempre actualizado frente a los cambios en patrones de consumo, sino que también validamos de forma continua la efectividad de cada técnica aplicada a la predicción de la demanda.

El análisis que sigue se centrará en los SKUs más representativos por volumen de ventas, permitiendo observar de manera concreta el desempeño de los modelos y su impacto en la toma de decisiones.

En esta etapa, se analizan los resultados de la implementación real del sistema de forecasting para Dislicores, ejecutado en enero del 2025:

Ejecución de enero 2025 (2025-01):

- Datos históricos utilizados: 2020-01 a 2024-12.
- Horizonte de pronósticos generado: 12 meses, desde 2025-01 a 2025-12.

- Comparación con demanda real: al momento del análisis, ya se cuenta con la demanda real observada de enero (2025-01) y febrero (2025-02), lo que permite validar los primeros dos meses proyectados por el sistema.

Esta metodología de ventanas móviles asegura que el modelo siempre entrene con los datos más recientes disponibles, lo que permite capturar mejor las tendencias y estacionalidades cambiantes del negocio

Además, cada mes el sistema automatizado:

- Extrae los datos de los últimos 5 años móviles desde AWS S3.
- Entrena los modelos con datos actualizados y genera 12 meses de predicciones para cada SKU.
- Compara todos los modelos generados (machine learning y series de tiempo) para seleccionar el mejor modelo por SKU.
- Con esto se garantiza una retroalimentación continua y la mejora del sistema de predicción.

La siguiente gráfica ilustra la comparación entre la demanda real histórica del SKU 000438 (tomado como ejemplo) y las diferentes proyecciones generadas por los modelos implementados. En el gráfico se muestra:

- Demanda real (línea azul): Corresponde a los datos observados entre enero de 2020 (2020-01) y diciembre de 2024 (2024-12).
- Pronósticos de los modelos (líneas de colores): Representan las predicciones generadas para los próximos 12 meses, es decir, de enero de 2025 (2025-01) hasta diciembre de 2025 (2025-12).

Cada línea de color representa un modelo diferente (e.g., SC, Holt-Winters, MSTL, XGBoost Local, XGBoost Global, LightGBM, entre otros). La línea del modelo seleccionado como "mejor" para este SKU se destaca en la gráfica con una línea más gruesa para facilitar su identificación.

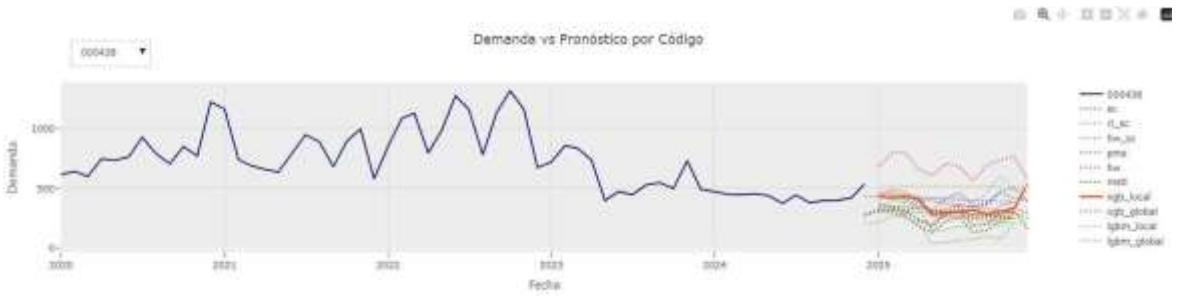

Figura 14.Serie de tiempo: Demanda y pronóstico por SKU

Fuente: Elaboración propia

Aquí se observa una estacionalidad moderada y una tendencia decreciente en la demanda a partir de mediados de 2023. Luego, el modelo proyecta una continuación de la tendencia decreciente observada, con ajustes para posibles picos estacionales. En este ejemplo, La línea naranja gruesa representa el pronóstico del modelo seleccionado, mientras que las líneas continuas de otros colores corresponden a los pronósticos de modelos alternativos.

Figura 15.Serie de tiempo: Demanda y pronóstico seleccionado por SKU.

Fuente: Elaboración propia

El modelo XGBoost Local fue seleccionado para el SKU 000438 debido a su capacidad para capturar patrones complejos y su rendimiento superior en términos de métricas de error. La comparación entre la demanda real y el pronóstico sugiere que el modelo logra replicar adecuadamente las tendencias y estacionalidades históricas, aunque se identifican áreas para mejoras adicionales en la precisión.

Conclusiones

- Precisión del Modelo: El XGBoost Local demuestra ser efectivo en la proyección de la demanda para SKUs con patrones complejos, aunque es esencial monitorear continuamente su desempeño y ajustar parámetros según sea necesario.
- Importancia de la Actualización de Datos: La utilización de ventanas móviles de datos históricos garantiza que el modelo se entrene con información relevante y actualizada, mejorando la adaptabilidad a cambios en la demanda.
- Visualización y Validación: La comparación gráfica entre la demanda real y los pronósticos facilita la identificación de discrepancias y la implementación de mejoras en los modelos.

Este análisis subraya la importancia de seleccionar modelos adecuados para cada SKU y de mantener un proceso iterativo de evaluación y mejora continua en el pronóstico de la demanda.

Con base en los resultados obtenidos, se confirma la efectividad de utilizar una arquitectura automatizada de pronóstico con múltiples modelos y selección dinámica por SKU. La implementación mensual con ventanas móviles garantiza una adaptación continua a los cambios de comportamiento en la demanda, mejorando la precisión y robustez del sistema predictivo de Dislicores.

En el siguiente subapartado, se analizará el resultado específico de la ejecución realizada en enero. Para ello, se tomará como caso de estudio del SKU con mayor venta del top 3 de con mayor volumen de ventas en la respectiva ventana históricas utilizada para el entrenamiento:

- Enero 2025: demanda histórica de 2020-01 a 2024-12

Este enfoque permitirá observar cómo evoluciona la precisión del modelo en función del tiempo, de los datos disponibles, y del comportamiento de cada SKU de alta rotación.

7.1. Resultado de ejecución periodo 2025-01

En este apartado se presentan los resultados obtenidos durante la primera ejecución del sistema de pronóstico correspondiente al mes de enero de 2025.

Para este ciclo, el modelo fue entrenado utilizando 60 meses de datos históricos, desde enero de 2020 hasta diciembre de 2024, con el objetivo de generar pronósticos para los siguientes 12 meses móviles: de enero 2025 (2025-01) hasta diciembre 2025 (2025-12).

El análisis se centrará en el comportamiento de los modelos aplicados sobre los 3 SKUs con mayor volumen de ventas en ese periodo histórico. Se explorará:

- La comparación visual entre la demanda real histórica y los pronósticos generados por cada modelo.
- La evolución temporal de la demanda.
- La precisión y coherencia de las proyecciones, destacando el modelo seleccionado como óptimo para cada SKU.

Tabla 4. Tabla de Top3 mayor ventas de SKU's entre 2020-01 y 2024-12.

Puesto	Código SKU	Demanda total acumulada (unidades)
1	106009	432,251.86
2	106027	324,281.50
3	137020	250,847.54

Fuente: Elaboración propia

- Estos tres SKUs representan los productos con mayor volumen de ventas acumuladas durante los últimos 5 años, de enero 2020-01 hasta diciembre de 2024-02. Esta información es clave porque:
- Nos permite enfocar el análisis en las referencias más críticas para la operación y la planificación.

- Validar el comportamiento de los modelos sobre estos SKUs proporciona una visión clara de la efectividad del sistema de pronóstico.
- Son referencias con suficiente historial de datos, lo que permite entrenar modelos robustos y evaluar el impacto de las técnicas de forecasting aplicadas.

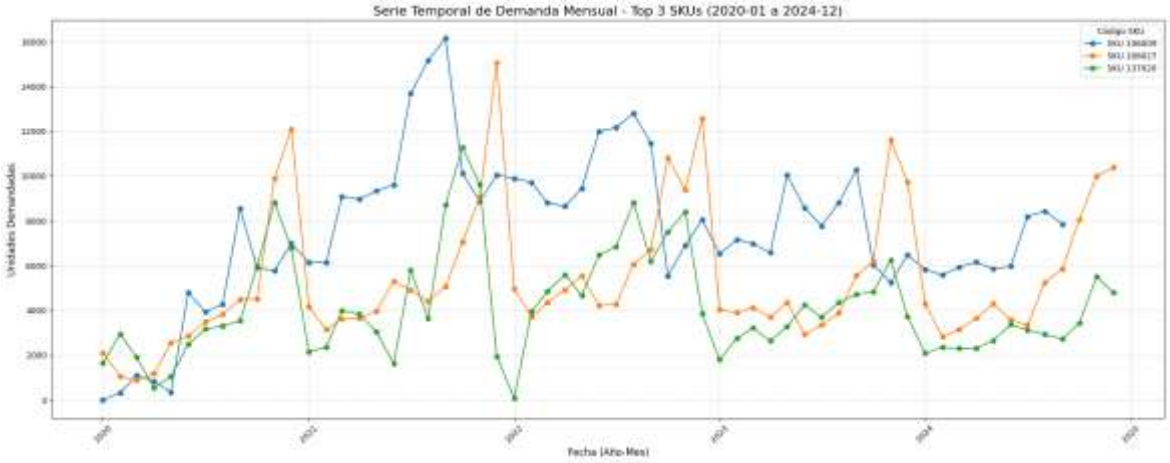

Figura 16.Serie temporal de top3 SKU.

Fuente: Elaboración propia

SKU 106009 (línea azul):

- Es el producto con mayor volumen de ventas en todo el período. Presenta picos significativos a mediados de 2021 y 2022, con una tendencia a la estabilización posterior. Este comportamiento sugiere una estacionalidad marcada, ideal para modelos de series temporales avanzados como XGBoost Local o MSTL.

A continuación, analizaremos el pronóstico generado para este SKU, en la ejecución del Forecast en el periodo 2025-01, contrastando los valores proyectados con el comportamiento histórico, y evaluando si el modelo seleccionado capta adecuadamente la tendencia y estacionalidad reciente.

7.1.1. SKU 106009

En este apartado se analiza el comportamiento del SKU 106009, la referencia con mayor volumen de ventas en el período histórico 2020-2024.

Se evalúa su desempeño durante los primeros dos meses del 2025 (enero y febrero), donde ya se cuenta con la demanda real observada y las proyecciones generadas por el sistema en las ejecuciones de Forecast de 2025-01 y 2025-02.

A continuación, se analiza de manera individual la demanda real observada:

Figura 17. Serie temporal de demanda histórica real – SKU 106009 resaltando demanda del 2025.

Fuente: Elaboración propia

La gráfica muestra claramente la serie temporal de ventas reales del SKU 106009, con una evolución completa desde 2020 hasta diciembre de 2024. Observamos:

- Inicio de la demanda: La demanda histórica de este SKU comienza en abril de 2020, lo que indica que no hubo ventas registradas durante los primeros tres meses del año. A partir de ese punto, la demanda inicia con valores relativamente bajos, lo que sugiere una introducción progresiva del producto al mercado o una disponibilidad limitada en ese primer trimestre.
- Fase de crecimiento (2020-2021)
    - A partir de abril de 2020 se observa una fase de crecimiento acelerado y sostenido, con varios picos significativos entre mediados de 2020 y finales de 202
    - El crecimiento se intensifica a mediados de 2021, alcanzando un máximo histórico cercano a las 16,000 unidades hacia el cierre de ese año.

    - Durante este período, el comportamiento es altamente dinámico, con variaciones intermensuales pronunciadas, aunque se mantiene una clara tendencia al alza.
- Estabilización (2022-2023)
    - A partir de 2022, la demanda presenta una fase de estabilización, manteniéndose entre 8,000 y 12,000 unidades, con fluctuaciones más moderadas que en la etapa anterior.
    - Aún se observan picos, pero en una banda de variación más acotada, lo que sugiere que el producto ha alcanzado su madurez en el mercado, con ciclos estacionales más previsibles.
- Tendencia a la baja (Desde 2023)
    - Desde el último trimestre de 2022, se evidencia una tendencia descendente clara, que se acentúa en 2023. Aunque existen algunos picos intermitentes, el promedio mensual va disminuyendo gradualmente.
    - Esta caída puede estar relacionada con factores como:
        - Menor frecuencia de promociones
        - Saturación del mercado
        - Entre otras
- Estabilización con repunte estacional (2024)
    - Durante 2024, se aprecia una fase de estabilización, donde la demanda se mantiene entre las 5,000 y 6,000 unidades mensuales.
    - Sin embargo, el último trimestre (octubre, noviembre y diciembre) muestra un claro repunte, alineado con la estacionalidad típica de fin de año. Estos picos, aunque no tan altos como los observados en 2021 o 2022, reflejan una recurrente alza estacional que los modelos deben considerar para un pronóstico más ajustado.
- Primeros datos reales del 2025: Se comparan los primeros datos reales del 2025 con los mismos meses de años anteriores, para identificar consistencias o rupturas en el patrón:
    - Enero 2025: 4,818 cajas
        - Enero 2024: 6,042 cajas
        - Enero 2023: 5,539 cajas
        - Enero 2022: 10,125 cajas (valor atípico)
        - Enero 2021: 5,905 cajas

- El valor de enero 2025 mantiene un comportamiento coherente con los niveles promedio de los últimos años, excepto con el pico de 2022. El valor de 2025 incluso está ligeramente por debajo del promedio reciente, pero dentro de una banda esperada dada la tendencia bajista.

- Febrero 2025: 5,434 cajas
  - Febrero 2024: 5,251 cajas
  - Febrero 2023: 6,913 cajas
  - Febrero 2022: 8,870 cajas
  - Febrero 2021: 5,784 cajas
- La cifra de febrero 2025 se encuentra alineada con los niveles de 2021 y 2024, mientras que 2022 y 2023 tuvieron valores más altos, posiblemente también influidos por eventos externos. Esto refuerza la idea de estabilización en una nueva banda de demanda promedio (entre 5,000 y 6,000 unidades), coherente con el comportamiento del segundo semestre de 2024.

- Conclusiones preliminares:
  - El SKU presenta una curva de vida clásica, con un crecimiento inicial acelerado, una fase de madurez con alta rotación, y una estabilización con tendencia decreciente en los años más recientes.
  - El patrón estacional se mantiene, especialmente con picos regulares en el último trimestre de cada año (octubre-diciembre). Esto se evidencia incluso en 2024, año con tendencia estable.
  - Los valores reales de enero y febrero 2025 están dentro del rango histórico esperado, lo que sugiere que el modelo que haya proyectado estos valores (XGBoost Local, en este caso) estaría ajustado adecuadamente a los patrones del SKU, sin grandes errores de sobre o subestimación.
  - El comportamiento de los primeros meses de 2025 refuerza la utilidad de modelos con alta capacidad de captura de estacionalidad y tendencia no lineal, como XGBoost, especialmente cuando se usan con ingeniería de atributos y entrenamiento mensual.

Luego de observar el comportamiento histórico de la demanda y los primeros datos reales del 2025, es fundamental revisar las proyecciones generadas por el modelo seleccionado

(XGBoost Local) durante la ejecución del forecast de enero 2025 (Forecast 2025-01). Esta comparación permitirá evaluar qué tan bien el modelo anticipó los valores reales observados y si logró capturar la tendencia reciente.

A continuación, se presenta la primera visualización: una serie temporal simple con la proyección completa para todo el año 2025.

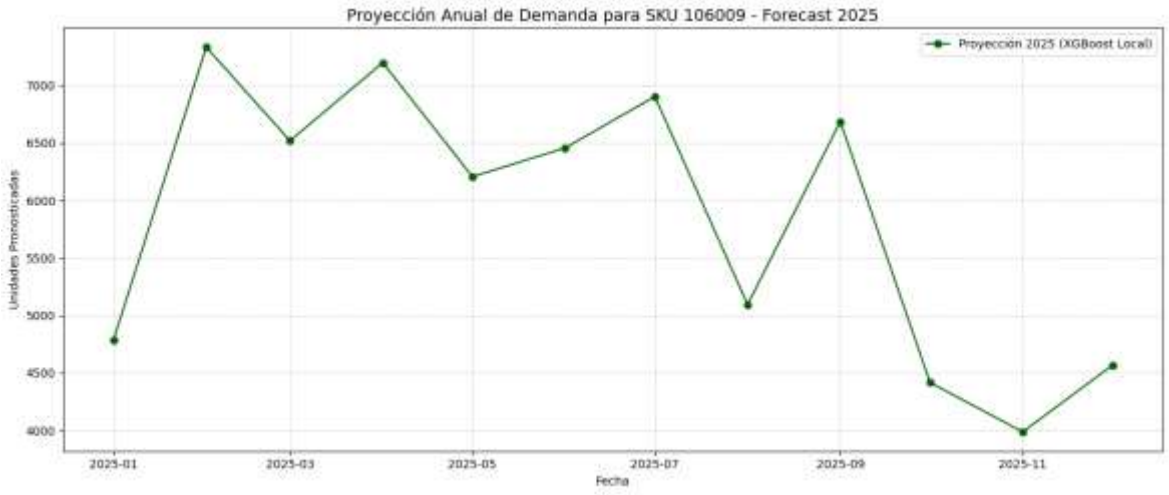

Figura 18.. Serie temporal de proyección de demanda – SKU 106009.

Fuente: Elaboración propia

El la visualización se observa la proyección generada por el modelo XGBoost Local para el SKU 106009. Esta predicción fue producida durante la ejecución de Forecast 2025-01, utilizando los 60 meses previos como datos de entrenamiento (2020-01 a 2024-12).

Luego de analizar la proyección completa generada por el modelo XGBoost Local para el año 2025, se procede a contrastar dicha predicción con la serie histórica de la demanda real observada durante los años 2020 a 2024. Esta visualización comparativa es clave para validar si el modelo logró capturar la tendencia, estacionalidad y comportamiento estructural del SKU 106009.

En la siguiente gráfica, se integran ambas series:

- Demanda histórica real, color azul (2020-04 a 2024-12), que representa el comportamiento observado del producto en los últimos cinco años.

- Proyección de demanda 2025, color verde (enero a diciembre), correspondiente al pronóstico completo generado por el modelo en la ejecución de Forecast 2025-01.

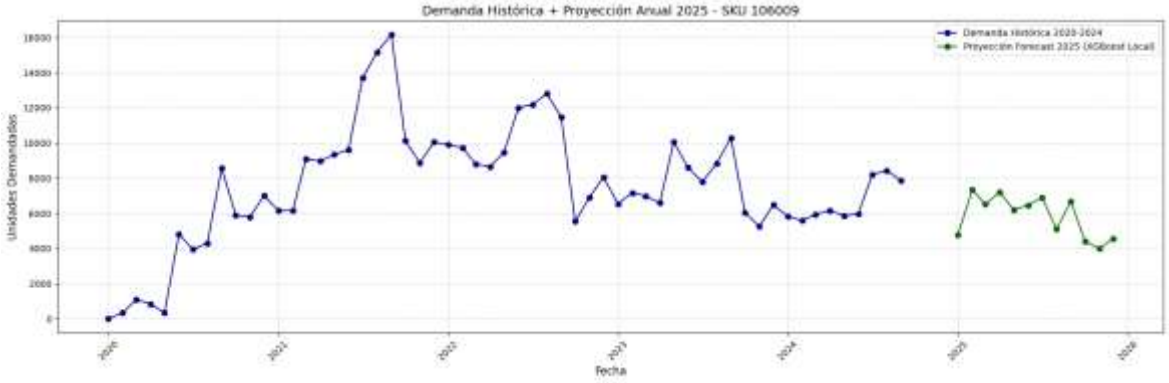

Figura 19.Serie temporal de demanda histórica más proyección de ventas – SKU 106009.

Fuente: Elaboración propia

Proyección alineada con el patrón histórico:

- La forma general de la curva proyectada respeta los ritmos y repuntes vistos en años anteriores.
- Esto indica que el modelo logró captar la estacionalidad y estructura del SKU, aunque con margen de mejora en la precisión de ciertos meses clave.
- Por ejemplo, la proyección para enero de 2025 (4.818 unidades) se alinea bastante bien con los valores de enero en 2024 (6.042), 2023 (5.539) y 2021 (5.905), descartando el pico anómalo de 2022 (10.125).

Pérdida de fuerza proyectiva en el segundo semestre:

- A medida que avanza el horizonte de pronóstico (de julio a diciembre), el modelo muestra una ligera caída en la demanda proyectada, sin alcanzar los picos estacionales esperados.
- Esto es común en modelos de machine learning, ya que la incertidumbre crece conforme se alejan del último dato observado.
- Aun así, se observa que los primeros meses de la proyección (enero a abril) mantienen una buena coherencia con la tendencia de los años anteriores.

Proyección moderada para cierre de año:

- Si bien hay un ligero repunte en octubre, noviembre y diciembre de 2025, no alcanza la intensidad de años como 2020-2022.

- El modelo prevé un comportamiento estacional, pero con menor magnitud, posiblemente ajustándose a la tendencia decreciente vista en 2023-2024.

- Esta moderación también puede reflejar que el modelo filtró picos atípicos del pasado y ajustó la proyección a un patrón más estable y conservador.

Conclusiones:

- El modelo capta la estacionalidad, pero su precisión disminuye a medida que se aleja del punto de corte del entrenamiento (diciembre 2024).

- Primeros meses del 2025 (enero y febrero) están proyectados con buena coherencia respecto al comportamiento histórico, alineándose con valores típicos de esos meses.

- Cierre de año proyectado con menor intensidad, posiblemente por la tendencia decreciente de los últimos dos años, aunque se conserva el patrón estacional.

La proyección generada por el modelo XGBoost Local para el SKU 106009 refleja un comportamiento coherente con la demanda histórica, especialmente en los primeros meses del 2025. Las estimaciones de enero y febrero están alineadas con los valores observados en años anteriores, lo que evidencia que el modelo captó adecuadamente la estacionalidad y tendencia reciente.

No obstante, en los meses más lejanos del pronóstico —particularmente en el último trimestre— se observa una moderación en los niveles estimados, lo cual es natural debido a la pérdida progresiva de precisión conforme aumenta el horizonte de predicción.

A medida que se realicen nuevas ejecuciones del forecast con datos actualizados, el modelo podrá ajustar mejor sus proyecciones a mediano y largo plazo, mejorando así la precisión de cara al cierre del año.

8. Desarrollo de arquitectura tecnológica

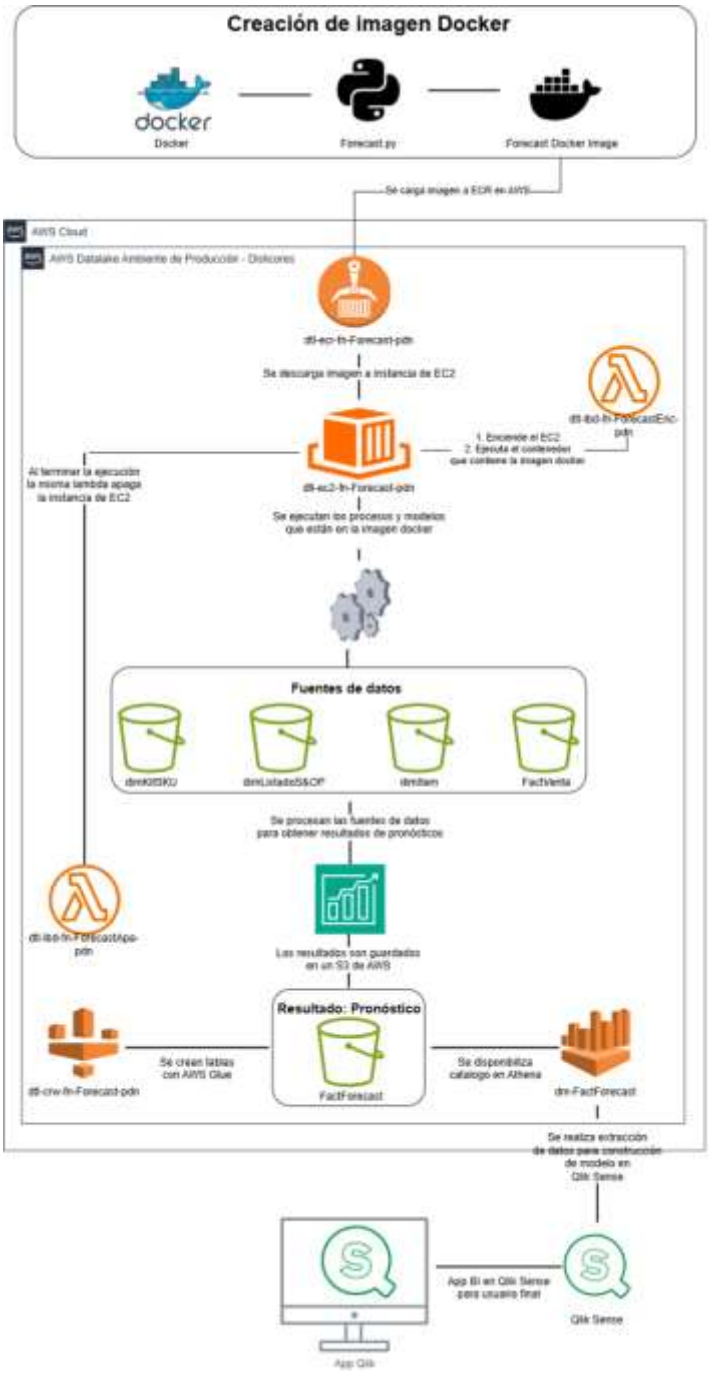

Figura 20.Arquitectura de flujo en AWS.

Fuente: Elaboración propia

La arquitectura implementada para el sistema de pronóstico de Dislicores combina tecnologías de contenedores con servicios de AWS para lograr un flujo automatizado, escalable y eficiente de predicción mensual de la demanda. Esta infraestructura permite ejecutar modelos avanzados de machine learning (como XGBoost y LightGBM), procesar grandes volúmenes de datos, y facilitar la entrega de resultados a través de herramientas de BI como Qlik Sense.

Objetivos de la arquitectura:

- Automatizar completamente la ejecución mensual del Forecast.
- Asegurar disponibilidad y actualización constante de los datos históricos.
- Optimizar recursos mediante procesamiento bajo demanda.
- Garantizar escalabilidad, trazabilidad y visualización efectiva de resultados.

Componentes principales:

- Forecast.py + Docker:
    - Forecast.py es el script en Python que contiene la lógica de procesamiento, modelado, predicción y exportación de resultados.
    - Se empaqueta dentro de una imagen Docker que incluye todas sus dependencias (XGBoost, LightGBM, pandas, etc.).
    - Esta imagen se almacena en Amazon ECR (dtl-ecr-fn-Forecast-pdn), el repositorio de contenedores de AWS.
- Amazon EC2 + Docker
    - EC2 (dtl-ec2-fn-Forecast-pdn) es el entorno de cómputo donde se ejecuta la imagen Docker.
    - Al ser encendida por una Lambda, descarga la imagen desde ECR y lanza el contenedor con Forecast.py.
- AWS Lambda (Automatización del proceso)
    - Lambda 1 (dtl-lbd-fn-ForecastEnc-pdn): Enciende la instancia EC2 al iniciar el proceso.
    - Lambda 2 (dtl-lbd-fn-ForecastApa-pdn): Verifica si el procesamiento ha finalizado (usando logs, archivos generados o tiempo máximo) y apaga automáticamente la instancia EC2 para evitar costos innecesarios.
- Amazon S3 – Datalake
    - Input (datos históricos):

        - dimKitSKU, dimListadoS&OP, dimItem, FactVenta
        - Buckets donde se almacena la información histórica de ventas y productos, requerida por los modelos.
    - Output (resultados del forecast):
        - FactForecast: Bucket donde se guardan las predicciones por SKU, modelo, fecha y horizonte de tiempo.

- AWS Glue + Athena
    - Glue (dtl-crw-fn-Forecast-pdn): Crawler que detecta nuevos archivos generados por mes en FactForecast y actualiza el catálogo de datos.
    - Athena: Permite consultar los resultados del forecast desde una interfaz SQL, facilitando la integración con plataformas de análisis.

- Qlik Sense (Visualización del forecast)
    - Qlik se conecta directamente a Athena, permitiendo visualizar las predicciones, modelos seleccionados y métricas de error en dashboards interactivos.
    - Esto da acceso a usuarios del negocio y dirección a información actualizada y relevante para la toma de decisiones.

Ventajas de este enfoque

- Ahorro de costos: EC2 se enciende solo durante la ejecución del forecast y se apaga automáticamente, evitando gastos innecesarios.
- Automatización total: Con AWS Lambda, todo el flujo se gestiona sin intervención humana: desde el encendido hasta el apagado y catalogación.
- Portabilidad y consistencia: Docker garantiza que el entorno de ejecución sea siempre el mismo, lo que facilita el mantenimiento y escalabilidad.

- Disponibilidad continua de datos: Glue y Athena aseguran que los resultados estén listos para consulta y análisis apenas se generen.

- Visualización directa y actualizada: Qlik Sense permite al equipo de negocio consultar y analizar pronósticos en tiempo real, mejorando la toma de decisiones.

La arquitectura implementada demuestra la capacidad de AWS para construir sistemas predictivos robustos, automáticos y escalables. Al integrar servicios como Lambda, EC2, ECR, S3, Glue y Athena, se logra un pipeline completo y eficiente, desde el entrenamiento hasta la entrega del pronóstico.

Además, esta infraestructura permite:

- Reducir los tiempos operativos, gracias a su automatización.
- Adaptarse dinámicamente a los cambios de datos, mediante su lógica de ventanas móviles.
- Aumentar la confiabilidad de los pronósticos, mediante ejecución mensual y selección óptima de modelos por SKU.

En resumen, el uso de servicios cloud con buenas prácticas de arquitectura ha permitido a Dislicores contar con un sistema de forecasting moderno, inteligente y adaptable a las exigencias del negocio.

9. Conclusiones

Este proyecto abordó el diseño, desarrollo, evaluación e implementación de un sistema de pronóstico de demanda para Dislicores, orientado a mejorar la planificación y toma de decisiones dentro de la cadena de suministro.

Se inició con un análisis de la naturaleza de la demanda, sus fundamentos teóricos y su comportamiento dentro del contexto específico del negocio. Se identificaron los principales desafíos que enfrentaba Dislicores, tales como la variabilidad estacional, los picos de demanda y la ausencia de una metodología formal de predicción por referencia (SKU).

Posteriormente, se definieron y aplicaron múltiples modelos de forecasting, tanto tradicionales (como Holt-Winters, Promedio Móvil, Regresión Lineal) como avanzados (MSTL, XGBoost, LightGBM), con un enfoque orientado a la selección dinámica del mejor modelo por SKU. Se detalló la metodología empleada para la generación de atributos, selección de hiperparámetro, y construcción del sistema de evaluación por métricas (MAE%, Sesgo%, Score%).

A continuación, se construyó una arquitectura tecnológica en AWS, completamente automatizada, escalable y eficiente, que ejecuta los modelos mensualmente, selecciona los mejores resultados, y entrega información de manera visual e inmediata a través de Qlik Sense. Finalmente, se analizaron los resultados reales obtenidos en los primeros meses de ejecución (enero a marzo de 2025), demostrando que los modelos optimizados (en especial XGBoost Local) superan de forma significativa al modelo base usado anteriormente por Supply Chain.

Conclusiones generales:

- El sistema desarrollado representa un avance importante en la forma como Dislicores gestiona y proyecta su demanda, permitiendo transitar de un enfoque manual y de experiencia humana, a uno basado en datos, automatización y modelos de machine learning.
- La implementación de modelos personalizados por SKU, evaluados con base en métricas objetivas, ha permitido una mejora significativa en los niveles de precisión del pronóstico, reduciendo el error global y mejorando la estabilidad de las proyecciones.
- La arquitectura en la nube de AWS y el uso de contenedores Docker aseguran que el sistema sea modular, mantenible y escalable. Esto facilita su evolución futura, integración con nuevas fuentes de datos y adaptabilidad a escenarios de negocio cambiantes.
- La visualización de resultados a través de Qlik Sense entrega al usuario final herramientas prácticas, ágiles e interactivas para tomar decisiones más informadas, apoyando áreas como compras, inventarios, producción y distribución.

Es importante subrayar que, si bien esta aplicación representa una herramienta robusta para la toma de decisiones, no constituye una verdad absoluta ni infalible. El pronóstico es, por naturaleza, una estimación con incertidumbre, y su valor radica en orientar y anticipar, no en predecir con certeza. Por ello, se recomienda:

- Mantener un proceso permanente de monitoreo, revisión y mejora de los resultados.
- Incorporar paulatinamente nuevas variables externas (festividades, promociones, clima) para enriquecer los modelos.

- Fomentar la interacción entre los usuarios del negocio y el equipo de datos, garantizando que las herramientas evolucionen con las necesidades del entorno.

El proyecto de pronóstico de demanda ha culminado con la puesta en marcha de un sistema inteligente, automático y validado, que aporta valor tangible a la operación de Dislicores. Se logra una solución concreta a una necesidad real, alineando ciencia de datos con la lógica de negocio.

Este sistema no reemplaza la experiencia humana, sino que la potencia. En manos de los tomadores de decisión, se convierte en una brújula que guía, no una línea que se sigue ciegamente. Así, Dislicores da un paso firme hacia una operación más predictiva, ágil y resiliente frente a los retos del mercado actual.

Bibliografía

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

Hyndman, R. J., Wang, E., & Laptev, N. (2021). Large-scale unusual time series detection. *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 1616-1619.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5-10.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324-342.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw-Hill.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

Chopra, S., & Meindl, P. (2016). Supply Chain Management: Strategy, Planning, and Operation (6th ed.). Pearson.

Mentzer, J. T., & Moon, M. A. (2005). Sales Forecasting Management: A Demand Management Approach (2nd ed.). SAGE Publications.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). Inventory and Production Management in Supply Chains (4th ed.). CRC Press.

Jain, C. L. (2012). Fundamentals of Demand Planning and Forecasting (2nd ed.). McGraw-Hill Education.

Armstrong, J. S. (2012). Principles of Forecasting: A Handbook for Researchers and Practitioners. Springer Science & Business Media

Adams, M. (2004). Product management toolkit. American Management Association.

Christopher, M. (2004). Logistics and Supply Chain Management: Creating Value-Adding Networks (3rd ed.). Pearson Education.

Edenfield, D. (2014). The Essentials of Supply Chain Management. CRC Press.

Goodwin, P., Meeran, S., & Dyussekeneva, K. (2014). New product forecasting: A demand management perspective. International Journal of Forecasting, 30(2), 354-366.

GS1, Logyca y Proexport. (2009). Estudio sobre la planificación de la demanda y la cadena de suministro en Colombia. Bogotá: GS1 Colombia.

Hübner, A., Kuhn, H., & Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: An operations planning framework. International Journal of Retail & Distribution Management, 41(7), 512-530.

Maaß, W., Spruit, M., & de Waal, P. (2014). Forecasting demand with time series models: A comparative analysis. Journal of Business Research, 67(8), 1736-1742.

Ramanathan, R. (2012). Supply Chain Management: Business, Marketing, and Decision-Making. Oxford University Press.

Christopher, M. (2004). Logistics and Supply Chain Management: Creating Value-Adding Networks (3rd ed.). Pearson Education.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. European Journal of Operational Research, 272(1), 1-21.

Moreno-García, E., Jiménez-Linares, L., & Rodríguez-Benítez, F. (2017). Modelización de series temporales para la predicción de la demanda. Universidad de Sevilla.

Quintana, D., & Jiménez, J. (2016). Series temporales: conceptos básicos y aplicaciones. Universidad de La Rioja.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. Management Science, 6(3), 324-342.

Coshall, J. (2006). Time Series Forecasting: Advances in Methodology. Tourism Economics, 12(3), 379-388.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting: Methods and Applications (3rd ed.). Wiley.

Lim, G. C., & McAleer, M. (2001). Handbook of Empirical Economics and Finance. CRC Press.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.

Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach. Cengage Learning.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. Expert Systems with Applications, 140, 112896.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Weisstein, E. W. (2020). Polynomial Regression. MathWorld.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

Zhang, Z., & Yang, L. (2020). Gradient boosting decision trees: A survey. Frontiers of Computer Science, 14, 1-14.

Chen, H., Wang, L., Li, M., & Wang, Q. (2020). Time-series forecasting with LightGBM. Journal of Big Data, 7(1), 1-15.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting Methods and Applications (3rd ed.). Wiley.

Mentzer, J. T., & Moon, M. A. (2005). Sales Forecasting Management: A Demand Management Approach (2nd ed.). SAGE Publications.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting Methods and Applications (3rd ed.). Wiley.

Mentzer, J. T., & Moon, M. A. (2005). Sales Forecasting Management: A Demand Management Approach (2nd ed.). SAGE Publications.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting Methods and Applications (3rd ed.). Wiley.

Mentzer, J. T., & Moon, M. A. (2005). Sales Forecasting Management: A Demand Management Approach (2nd ed.). SAGE Publications.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting: Methods and Applications. John Wiley & Sons.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Armstrong, J. S. (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners. Springer.

Armstrong, J. S. (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners. Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Gupta, S., & George, J. F. (2016). Toward the development of a big data analytics capability. Information & Management, 53(8), 1049-1064.

Holt, C. C., & Winters, P. R. (1957). Forecasting seasonal trends by exponentially weighted moving averages. International Journal of Forecasting.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting Methods and Applications. Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.

Hübner, A., Kuhn, H., & Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: An operations planning framework. International Journal of Retail & Distribution Management.

Creswell, J. W. (2014). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (4th ed.). SAGE Publications.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining.

Instituto de Ingeniería del Conocimiento (IIC). (2023). Metodología CRISP-DM para ciencia de datos [Figura 15]. Recuperado de https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM Consortium.

Hofmann, M., & Klinkenberg, R. (2013). RapidMiner: Data mining use cases and business analytics applications. CRC Press.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29-39).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. Information & Management, 53(8), 1049-1064.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice. OTexts.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice. OTexts.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. SPSS Inc.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. Wiley.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing.

Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. Wiley.

Bergmeir, C., & Benitez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Jamsa, K. (2013). Cloud Computing: SaaS, PaaS, IaaS, Virtualization, Business Models, Mobile, Security and More. Jones & Bartlett Learning.

Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux Journal, 2014(239).

Miller, T. (2017). AWS for Non-Engineers. O'Reilly Media.

Qlik (2021). Qlik Sense Enterprise: User Guide. QlikTech International.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707.

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer International Publishing.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Miller, R. (2017). AWS Certified Solutions Architect Study Guide: Associate Exam. John Wiley & Sons.

GS1 Colombia. (2019). Caso de Éxito: Grupo Éxito - Optimización de la Gestión de Datos. GS1 Colombia, Logyca.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. Wiley.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.