UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL
INGENIERÍA DE TECNOLOGÍAS DE LA INFORMACIÓN Y SISTEMAS

Aplicación de modelos de Machine Learning para la planificación de la demanda en la empresa CBC Peruana S.A.C

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para:

Obtener el título profesional de Ingeniero Industrial y Comercial,

Obtener el título profesional de Ingeniero de Tecnologías de la Información y Sistemas

AUTORES

Maciel Carpio, Zannie Xilena
Salas Barrera, Felipe Alvaro
Sanchez Anticona, Crishtian Sebastian
Sanchez Chacon, Gabriela de los Angeles
Santana Fernandez, Jose Daniel

ASESOR

Fabian Arteaga, Junior John

ORCID Nº0000-0001-9804-7795

Noviembre, 2023

Trabajo Final- Grupo 07 TSP 2023

INFORME DE ORIGINALIDAD

11%	10%	2%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	2%
2	Submitted to Universidad ESAN -- Escuela de Administración de Negocios para Graduados Trabajo del estudiante	1%
3	repositorio.esan.edu.pe Fuente de Internet	<1%
4	bibliotecadigital.udea.edu.co Fuente de Internet	<1%
5	repositorio.ug.edu.ec Fuente de Internet	<1%
6	es.scribd.com Fuente de Internet	<1%
7	pirhua.udep.edu.pe Fuente de Internet	<1%
8	tesis.pucp.edu.pe Fuente de Internet	<1%

Resumen

La industria de bebidas enfrenta desafíos específicos en la planificación de la demanda, ya que la variabilidad de los patrones de consumo y la imprevisibilidad del cliente exige a las empresas establecer estrategias para satisfacer la demanda. El presente trabajo de investigación se centra en la aplicación de técnicas de Machine Learning para pronosticar la demanda de dos productos clave de la empresa CBC Peruana S.A.C: paquetes de gaseosa Concordia de Piña de 03 litros de 04 unidades y paquetes de gaseosa Evervess Ginger de 1,5 litros de 06 unidades. Para ello, se utilizaron los modelos de Regresión lineal, LightGBM Regressor y series de tiempo, como SARIMA y FB Prophet, aplicando los enfoques de Forecasting y Regresión. La evaluación de modelos se realizó utilizando métricas como MAE, MAPE y RMSE. Entre los resultados obtenidos, se obtuvo que el modelo FB Prophet registra un MAPE promedio de 24.64, MAE promedio de 685.16 y un RMSE promedio de 1003.90.

Este estudio proporciona una base sólida para futuras investigaciones en la aplicación de Machine Learning en la industria de bebidas y demuestra el potencial de estas tecnologías para transformar las operaciones comerciales y mejorar la competitividad en el mercado.

Palabras clave: Machine Learning, demanda, Regresión lineal, LightGBM Regressor, SARIMA, FB Prophet, MAE, MAPE, RMSE

Abstract

The beverage industry faces specific challenges in demand planning, given the variability in consumption patterns and the unpredictability of customer behavior, requiring companies to establish strategies to meet demand. This research focuses on the application of Machine Learning techniques to forecast the demand for two key products of the company CBC Peruana S.A.C: 3-liter, 4-unit packs of Concordia Pineapple soda and 1.5-liter, 6-unit packs of Evervess Ginger soda. To achieve this, models such as Linear Regression, LightGBM Regressor, and time series models like SARIMA and FB Prophet were employed, applying Forecasting and Regression approaches. Model evaluation was conducted using metrics such as MAE, MAPE, and RMSE. Among the results obtained, the FB Prophet model recorded an average MAPE of 24.64, average MAE of 685.16, and average RMSE of 1003.90.

This study provides a solid foundation for future research on the application of Machine Learning in the beverage industry and demonstrates the potential of these technologies to transform business operations and enhance competitiveness in the market.

Key Word: Machine Learning, demand, Linear regression, LightGBM Regressor, SARIMA, FB Prophet, MAE, MAPE, RMSE

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Introducción

En el dinámico mundo empresarial, la capacidad para anticipar y gestionar la demanda de productos se ha convertido en un desafío crucial para las organizaciones. La precisión en las predicciones de demanda no solo impacta la eficiencia operativa, sino que también influye directamente en la satisfacción del cliente y, en última instancia, en la rentabilidad de la empresa. Las fluctuaciones en la demanda pueden generar problemas significativos como el exceso de inventario o la falta de productos en stock, afectando tanto las finanzas como la reputación de la empresa.

En este contexto, técnicas avanzadas de análisis de datos, particularmente Machine Learning, han surgido como herramientas poderosas para prever patrones y tendencias en los datos históricos, permitiendo así a las empresas tomar decisiones informadas y estratégicas.

En este sentido, la presente investigación, tiene como objetivo desarrollar y evaluar modelos de Machine Learning bajo el enfoque de Forecasting y Regresión que permitan predecir con precisión la demanda de dos productos de la empresa CBC Peruana S.A.C: paquetes de gaseosa Concordia de Piña de 03 litros de 04 unidades y paquetes de gaseosa Evervess Ginger de 1,5 litros de 06 unidades, los cuales representan variaciones significativas en la planificación de la demanda. Para ello, bajo el enfoque de Regresión se construyeron los modelos de Regresión Lineal Supervisada y LightGBM Regressor, y para el enfoque de Forecasting se adaptaron modelos de Regresión Lineal Supervisada, LightGBM Regressor, SARIMA y FB Prophet, mediante la aplicación de diferentes frecuencias de datos y técnicas de normalización. Para la evaluación de los modelos, se utilizará métricas clave como Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE) y Mean Absolute Error (MAE), con el objetivo de identificar el modelo que mejor se ajuste a las necesidades específicas de la empresa CBC Peruana S.A.C.

Del mismo modo, la implementación del modelo más efectivo representa un impacto significativo hacia la mejora de la capacidad de producción de la empresa CBC Peruana S.A.C. Asimismo, pretende contribuir en la gestión efectiva del inventario y en la toma de decisiones estratégicas, basadas en datos sólidos y proyecciones precisas de la demanda.

La investigación, se estructura en seis capítulos. En el capítulo I, se aborda la situación problemática de la compañía CBC Peruana S.A.C., así como la justificación (desde el punto de vista teórico, práctico y metodológico) y los límites de la investigación. En el capítulo II, se expone el marco teórico, el cual contiene la revisión de antecedentes y bases teóricas respecto a las técnicas de Machine Learning que se aplicarán en la presente investigación. En capítulo III, se desarrolla el entorno empresarial de la compañía CBC Peruana S.A.C, que involucra la descripción de la empresa, objetivos estratégicos, el modelo de negocio, la evaluación interna y externa, y mapa de procesos actual. El capítulo IV, desarrolla la metodología empleada en la investigación, donde se describe el diseño, el enfoque, el alcance de la investigación, la definición de la población y muestra; asimismo, el cronograma de las actividades y el presupuesto. El capítulo V, aborda el desarrollo de la solución de la investigación, y se describe la propuesta y evaluación de la solución. Finalmente, capítulo VI, expone las conclusiones de los efectos logrados y las propuestas para investigaciones posteriores.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción de la Realidad Problemática

El mercado de bebidas abarca las bebidas alcohólicas y no alcohólicas, bebidas y otros líquidos potables destinados al consumo humano, incluyendo cerveza, vino, refrescos, jugos de frutas, agua envasada o embotellada, entre otros.

Según el informe de Mordor Intelligence (2022), la industria de bebidas se centra en los precios y en satisfacer las necesidades y estilos de vida en constante cambio de los clientes. Asimismo, refiere que, en la última década, la industria de bebidas ha experimentado avances significativos en términos de innovación y diversificación de productos. Las empresas están enfrentando los desafíos crecientes en el mercado al introducir nuevos sabores, considerando las preocupaciones de salud y bienestar de los consumidores y la intensa competencia en el mercado, liderada por actores clave como Nestlé SA, Pernod Ricard, Diageo PLC, PepsiCo, Inc. y The Coca-Cola Company. Varias empresas activas en el mercado han adoptado la innovación de productos como estrategia, debido a los cambios en las preferencias de los consumidores a nivel mundial.

Esta necesidad, exige a las empresas comprender las tendencias y patrones de los clientes. También ayuda a personalizar las experiencias, mejorar sus ofertas y controlar la calidad. A partir de ello, las empresas pueden identificar nuevas oportunidades y actuar en consecuencia antes que la competencia.

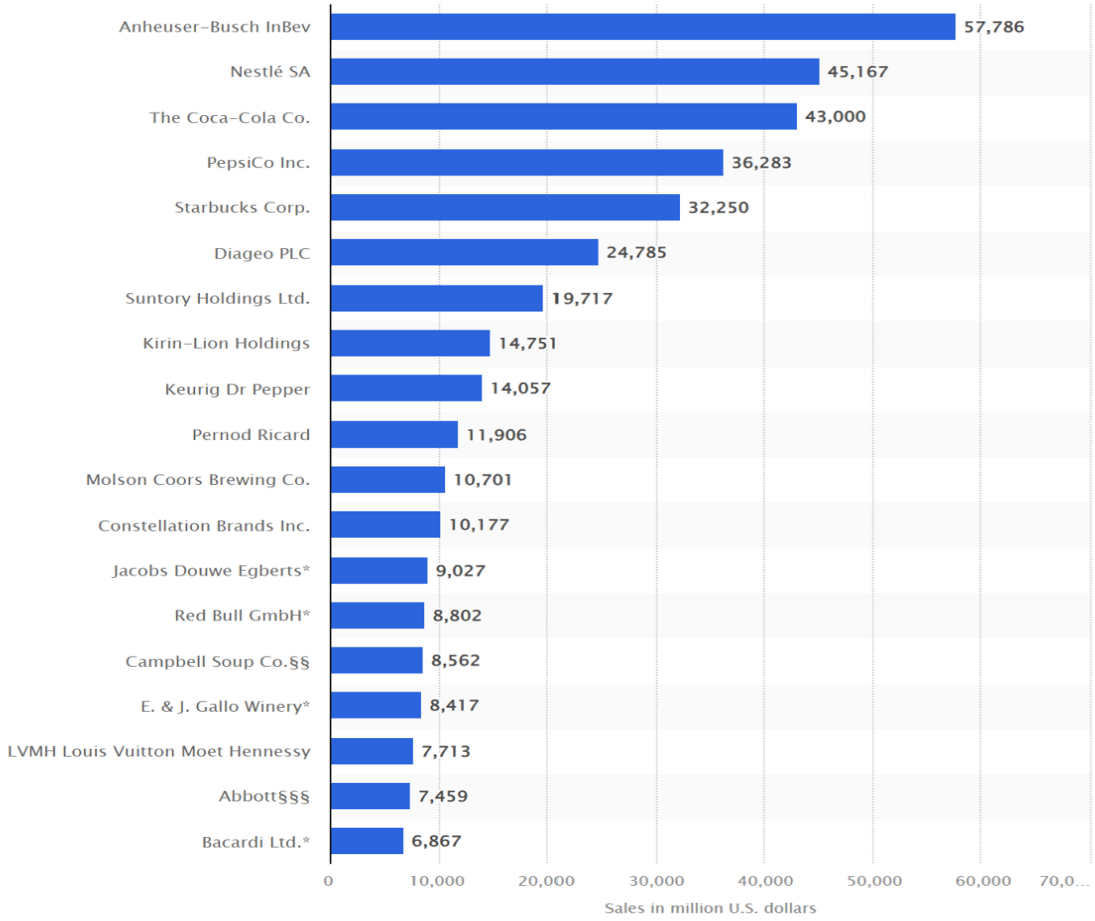
Para el caso de las bebidas carbonatadas, según Mordor Intelligence (2022), se proyecta que el mercado mundial registre una tasa de crecimiento de 6,5% en los próximos cinco años.

Según el portal de estadísticas globales Statista (2023), a nivel mundial se consumieron aproximadamente 272,5 mil millones de litros de bebidas alcohólicas, lo que representa una disminución en el consumo de más de 9,5 mil millones de litros en comparación con Estados Unidos, país que representa el mayor consumo de bebidas no alcohólicas a nivel mundial. En el año 2022, se consumieron en el mercado estadounidense más de 141,300 millones de litros de refrescos, zumos y agua embotellada. China y México ocuparon el segundo y tercer lugar, respectivamente.

Según Statista (2023), las empresas Anheuser – Bush InBev, Nestlé, The Coca Cola, PepsiCo y Starbucks Corporation, en el año 2022, obtuvieron mayor cuota de participación de mercado en la industria de bebidas.

Figura 1

Principales empresas de la industria de bebidas



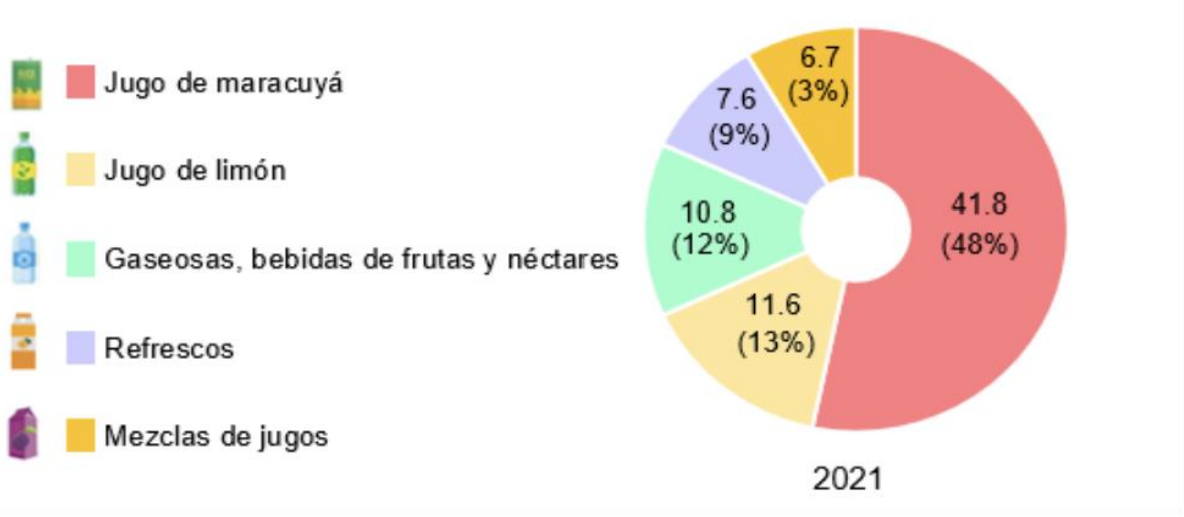
Nota. Adaptado de Ranking de las principales empresas de bebidas a nivel mundial en función de su facturación en 2022, por Abigail Orús, 2023.

Dada la variabilidad observada en la demanda de productos bebibles en el mercado peruano, resulta imperativo llevar a cabo actualizaciones y mejoras en las infraestructuras existentes, anticipándose proactivamente a los cambios y variaciones sustanciales en la demanda. El progreso tecnológico experimentado en los últimos años y la abundancia de datos almacenados han generado nuevas perspectivas para la exploración de métodos que permitan anticipar con mayor precisión las fluctuaciones en la demanda.

Según el reporte de la Asociación de Exportadores (2022), en el año 2021, en el Perú se alcanzó los US$ 87.4 millones en exportaciones de bebidas sin alcohol, estas incluyen jugos de limón, jugos de maracuyá, refrescos, gaseosas y otras mezclas de jugos. Con relación a las bebidas alcohólicas, refiere que, las exportaciones alcanzaron 39,6 millones de dólares al año 2022, lo que representa un crecimiento de 38,2% respecto al año 2021.

Figura 2

Exportaciones de principales bebidas



Nota. Adaptado de Reporte de tendencias, exportaciones de Perú de principales bebidas no alcohólicas, por ADEX, 2022.

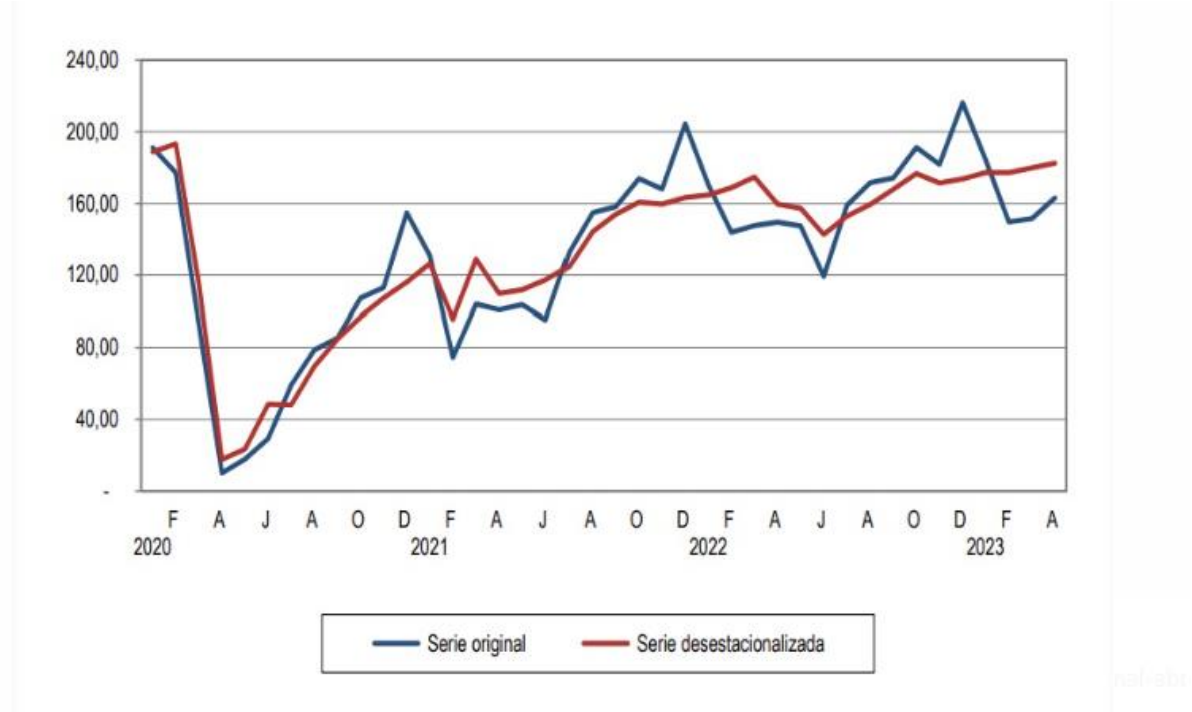
De acuerdo con el Instituto Nacional de Estadística e Informática (INEI), en el año 2021, se estima que el consumo per cápita de bebidas no alcohólicas (gaseosas) en el Perú es de, aproximadamente, 27 litros al año. Sin embargo, en las regiones de la costa, este consumo promedio se incrementa a alrededor de 30 litros. Adicionalmente, se menciona que el nivel socioeconómico tiene influencia en el consumo y que podría llegar a alcanzar un promedio de 47 litros.

De lo expuesto, el informe de Variación de Indicadores de Precios al Consumidor del Instituto Nacional de Estadística e Informática (INEI), en el año 2023 destaca que durante el 2022 se registraron cambios en los precios de las bebidas gasificadas, con una variación del 5,8%, así como en las aguas minerales y de mesa, con una variación del 5,2% en comparación con el periodo anterior. Estos cambios se atribuyen principalmente a factores externos, como el transporte.

Además, de acuerdo con el informe técnico de Producción Nacional del Instituto Nacional de Estadística e Informática (INEI), en el año 2023, se observó un aumento en la producción de bebidas sin alcohol durante el primer trimestre en comparación con el año anterior. Este crecimiento incluye bebidas como gaseosas, agua mineral, refrescos y bebidas hidratantes. El incremento se atribuye al crecimiento de negocios como juguerías, cafeterías, restaurantes y servicios de venta en línea, los cuales han experimentado un crecimiento continuo desde el mes enero del 2020 hasta abril del 2023.

Figura 3

Índice de la producción de alojamiento y restaurantes



Nota. Adaptado de Producción Nacional, por INEI, 2023.

La presente investigación se centra en la demanda de la empresa CBC Peruana S.A.C., que se dedica a colocar en el mercado productos como bebidas gaseosas y aguas minerales.

- Bebidas de PepsiCo: PepsiCo ofrece una amplia variedad de bebidas, siendo el portafolio más diversificado en el ámbito mundial.
- Bebidas funcionales de Beliv: Es una unidad de negocio dentro de la compañía. Estas bebidas han sido desarrolladas con la finalidad de ofrecer diversos beneficios específicos para la salud."

- Bebidas alcohólicas de Diageo: Destacada por la excelencia de sus licores y cervezas, las cuales están diseñadas con el propósito de atender la demanda de bebidas alcohólicas cumpliendo con elevados estándares industriales.

Actualmente, la Unidad de Producción y Operaciones de la empresa CBC Peruana S.A.C, es responsable de desarrollar la estrategia de previsión de la demanda, esto se realiza bajo la jefatura de Planificación logística. La planificación de la demanda se lleva a cabo para cada uno de los productos ofrecidos por la empresa, basado en ventas históricas, que posteriormente, son enviados a la Unidad de producción para su elaboración.

La empresa CBC Peruana S.A.C, opera con 02 plantas de producción, situadas en la provincia de Sullana, en la región de Piura y en Huachipa, Lima.

Tabla 1

Planta - Capacidad

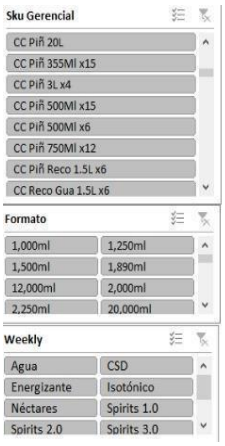
Planta	Capacidad de producción	Número de Línea
Sullana	1,3 millones de litros	2 líneas de envasado
Huachipa	3,2 millones de litros	3 líneas de envasado

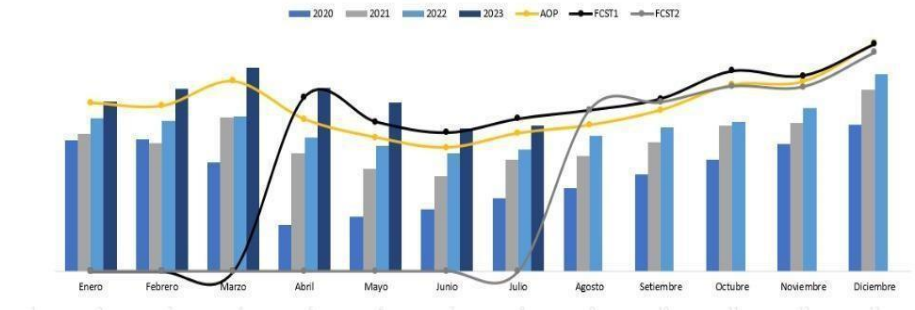
Nota. Elaboración propia

El volumen total de demanda de CBC Peruana S.A.C., se distribuye en 150 productos diferentes, entre bebidas alcohólicas, gaseosas, bebidas no carbonatadas. La cantidad media de paquetes de bebidas consumidas mensuales en el primer semestre del año 2023 se ha elevado a aproximadamente a 06 millones.

Figura 4

Demanda total de 150 productos





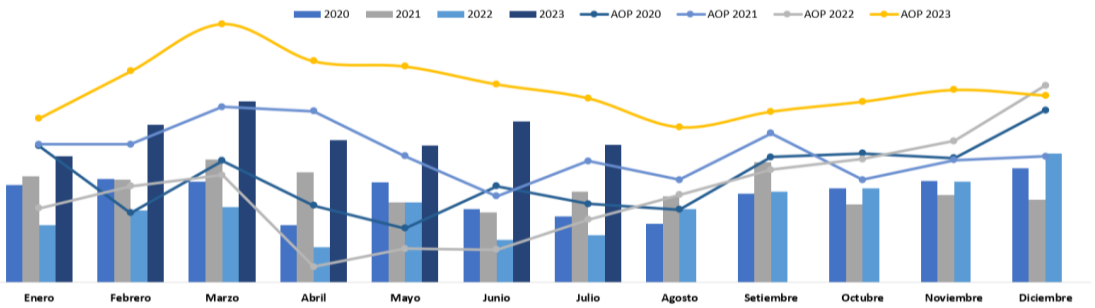
PAQ	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Setiembre	Octubre	Noviembre	Diciembre
2020	4,812,082	4,847,506	4,004,157	1,718,676	2,027,490	2,268,372	2,687,076	3,049,711	3,558,251	4,097,319	4,663,956	5,389,054
2021	5,040,319	4,705,347	5,642,356	4,347,610	3,752,323	3,483,897	4,090,036	4,247,350	4,735,553	5,343,554	5,448,832	6,656,901
2022	5,614,360	5,524,098	5,669,357	4,902,209	4,605,693	4,338,424	4,470,263	4,968,164	5,280,652	5,497,197	5,971,293	7,241,551
2023	6,214,150	6,679,976	7,477,356	6,733,680	6,189,913	5,241,436	5,332,197	-	-	-	-	-
AOP	6,181,802	6,073,064	6,976,293	5,578,689	4,908,916	4,534,806	5,067,982	5,361,566	5,917,287	6,844,057	6,972,415	8,378,395

Nota. Resultados de CBC Peruana S.A.C

De la revisión de la demanda total de la empresa CBC Peruana S.AC, se identificó que, la demanda real de bebidas Concordia de Piña de 03 litros de 04 unidades por paquete y Evervess Ginger de 1,5 litros de 06 unidades por paquete, presentan variaciones significativas respecto a la demanda pronosticada, entre los años 2020, 2021, 2022 hasta julio del año 2023.

Figura 5

Demanda de paquetes de bebida Evervess Ginger



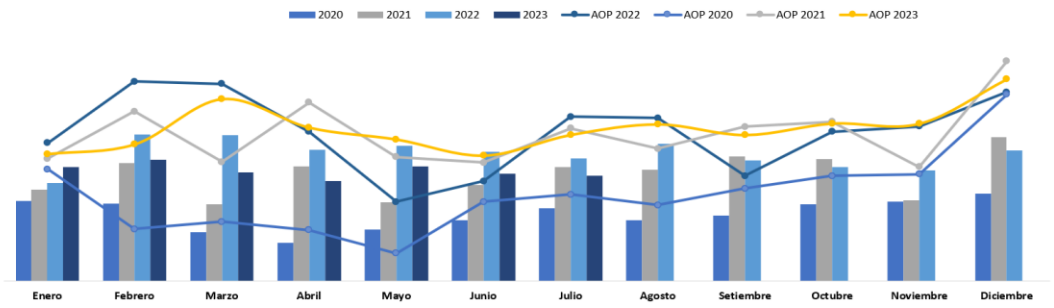
PAQ	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Setiembre	Octubre	Noviembre	Diciembre
2020	17,480	18,596	18,012	10,264	17,894	13,176	11,824	10,434	15,844	16,892	18,139	20,468
AOP 2020	24,471	12,459	21,794	13,754	9,663	17,260	14,070	13,042	22,498	23,142	22,310	30,906
2021	19,055	18,464	22,040	19,703	14,343	12,550	16,246	15,470	21,586	14,031	15,678	14,789
AOP 2021	24,771	24,742	31,517	30,736	22,661	15,436	21,769	18,409	26,766	18,380	21,949	22,627
2022	10,200	12,879	13,456	6,280	14,375	7,553	8,405	13,174	16,284	16,873	18,113	23,115
AOP 2022	13,260	17,258	19,242	2,763	6,037	5,816	11,262	15,677	20,192	22,104	25,358	35,366
2023	22,659	28,287	32,441	25,456	24,536	28,919	24,683	-	-	-	-	-
AOP 2023	29,457	37,904	46,391	39,711	38,768	35,570	33,075	27,890	30,678	32,456	34,567	33,567

MAE 2020	6,992	6,137	3,782	3,490	8,231	4,085	2,247	2,608	6,654	6,250	4,172	10,439
MAE 2021	5,716	6,278	9,477	11,034	8,319	2,886	5,524	2,939	5,181	4,350	6,271	7,838
MAE 2022	3,060	4,379	5,786	3,517	8,337	1,737	2,858	2,503	3,908	5,231	7,245	12,251
MAE 2023	6,798	9,617	13,950	14,255	14,231	6,651	8,392	27,890	30,678	32,456	34,567	33,567

Nota. Demanda de la bebida Evervess Ginger de 1,5 litros

Figura 6

Demanda de paquetes de bebida Concordia Piña



PAQ	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Setiembre	Octubre	Noviembre	Diciembre
2020	72,004	69,702	44,060	34,310	46,321	54,533	65,648	54,800	58,746	69,194	71,413	78,956
AOP 2020	100,805	46,700	53,312	45,976	25,013	71,438	78,121	68,500	83,420	94,795	96,448	168,630
2021	82,441	106,471	69,380	103,297	70,969	86,894	102,921	100,347	112,297	109,827	72,969	129,837
AOP 2021	110,471	153,319	107,539	161,144	112,131	106,879	137,914	119,413	139,248	143,874	102,886	198,651
2022	88,185	132,055	131,506	118,325	122,346	116,802	110,667	123,702	108,860	102,853	99,733	117,640
AOP 2022	124,641	180,154	178,053	134,987	71,385	89,938	148,294	147,205	94,987	134,738	139,626	170,589
2023	103,121	109,585	97,941	90,109	103,573	96,681	94,947	-	-	-	-	-
AOP 2023	114,711	123,515	164,375	138,565	127,865	113,282	131,984	141,531	131,886	142,272	142,155	182,298

MAE 2020	28,801	23,002	9,253	11,665	21,308	16,905	12,473	13,700	24,673	25,602	25,035	89,674
MAE 2021	28,030	46,847	38,159	57,846	41,162	19,986	34,993	19,066	26,951	34,046	29,917	68,814
MAE 2022	36,456	48,099	46,548	16,662	50,961	26,864	37,627	23,503	13,874	31,884	39,893	52,949
MAE 2023	11,590	13,929	66,434	48,456	24,292	16,601	37,037	-	-	-	-	-

Nota. Demanda de bebida Concordia de Piña de 03 litros

La fluctuación en la demanda ha generado un impacto en los costos de inventario, ya que la empresa CBC Peruana S.A.C, ha tenido que gestionar un exceso de inventario. Estos costos han sido significativos, alcanzando un total de S/7,036,815.84 en los primeros siete meses del 2023, para una demanda promedio de, aproximadamente, 3,9 millones.

Tabla 2

Costos mensuales del año 2023 (soles)

Costos	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio
Producción, Inventario y Transporte	51,150,1567	58,076,525	66,324,846	59,057,305	54,918,264	49,379,693	48,904,313
Exceso de Inventario	6,909,047	7,448,056	7,925,167	7,827,915	6,533,012	6,556,916	6,057,595
Total	**58,059,204**	**65,524,585**	**74,250,013**	**66,885,221**	**61,451,271**	**55,936,609**	**54,961,91**

Nota. Resultados de CBC Peruana S.A.C

Desde el año 2016 al año 2020, la empresa ha experimentado una disminución en la rotación de productos terminados. En el año 2020, el índice de rotación de inventarios fue negativo, esto, relacionado a factores externos, como la pandemia (COVID 19), que ha afectado la demanda y el flujo de inventario en el sector.

Tabla 3

Índice de Rotación de Inventarios

Periodos	2015	2016	2017	2018	2019	2020
Rotación de Inventarios	15.77	23.84	11.82	10.66	8.96	-0.99

Nota. Resultados de CBC Peruana S.A.C

Según, Mediavilla, M. A., Dietrich, F., & Palm, D. (2022), la mejora de los métodos de previsión de la demanda se ha convertido en una preocupación cada vez más importante para los distribuidores, fabricantes, y minoristas. Sin embargo, se presentan numerosos desafíos que deben afrontar las empresas al momento de proyectar su demanda. Entre ellos están: la carencia de información precisa, adecuada y oportuna para un pronóstico de demanda, la dificultad en identificación de causalidad entre variables correlacionadas, la presencia de desafíos idiosincráticos a cada mercado como alta volatilidad en las preferencias de los consumidores o productos perecibles y la dificultad en reducir el tiempo de pronosticación. Para superar estos desafíos, es necesario que las empresas en el rubro de bebidas y alimentos optimicen sus procesos de proyección de demanda. Esto implica desarrollar y aplicar herramientas de gestión de la demanda que sean avanzadas, considerando factores restrictivos con el objetivo de reducir al mínimo los errores humanos al manipular datos.

Existe una necesidad de optimización de los pronósticos y obtención de datos precisos que contribuyan a mejorar el flujo en la cadena de suministro.

En este contexto, esta investigación, radica en aplicación de técnicas de Machine Learning, que permitirá a la compañía mejorar la precisión de la previsión de la demanda, lo que a su vez favorecerá una óptima gestión de los recursos y una minimización de los costos vinculados con el excedente o la escasez de inventario. Además, el uso de estas técnicas permitirá una toma de decisiones más informada y eficiente, ya que se basará en datos precisos y actualizados. Asimismo, se provee que la aplicación de técnicas de Machine Learning

disminuya los riesgos relacionados con la gestión de la demanda y aumente la eficiencia financiera de la empresa.

1.2 Justificación de la Investigación

1.2.1 Teórica

El objetivo de la investigación radica en aplicar técnicas de Machine Learning, para pronosticar la demanda de acuerdo con las necesidades del cliente mediante la construcción de los modelos de Regresión Lineal, LightGBM Regressor y series temporales, como SARIMA y FB Prophet, desde el enfoque de Forecasting y Regresión.

Aamer, A., Eka Yani, L., & Alan Priyatna, I. (2020), señalan que, mientras más eficiente, transparente, resiliente y receptiva sea la cadena de suministro, mejores ingresos y ganancias podrá obtener la organización. La exactitud de la estimación de la demanda es un factor clave para la eficiencia de la gestión de la cadena de suministro, por lo que es necesario desarrollar modelos fiables de previsión de la demanda para realizar predicciones mejores y más precisas. Asimismo, precisan que, Machine Learning, es una herramienta disruptiva que puede utilizarse para desarrollar mejores modelos de previsión de la demanda que los que se utilizan actualmente en la administración de la cadena de suministro.

1.2.2 Metodológica

La investigación, se basará en la metodología Industry Standard Process for Data Mining, ampliamente reconocida en el análisis de datos.

La recopilación de datos comprenderá las ventas históricas desde enero del 2019 hasta julio del 2023 de dos productos específicos: paquetes de 04 unidades de gaseosa Concordia de Piña de 03 litros y paquetes de 06 unidades de gaseosa Evervess Ginger de 1,5 litros. A partir de estos datos, se construirán cuatro modelos de Machine Learning: Regresión Lineal, LightGBM Regressor y series temporales (SARIMA y FB Prophet).

1.2.3 Práctica

El presente estudio aplicará técnicas de Machine Learning para la previsión de la demanda en la industria de bebidas de gaseosa de la empresa CBC Peruana S.A.C, a fin de reducir la variación de los errores de la planificación de la demanda, y contribuya a:

- Planificación empresarial: Ayuda a la empresa a determinar los recursos necesarios para la producción y venta de sus productos.
- Establecimiento de metas y objetivos realistas: Los pronósticos de la demanda son útiles para establecer objetivos y metas realistas, y evaluar el desempeño empresarial en comparación con los pronósticos originales.
- Reducción de costos: Implica minimizar los costos asociados con situaciones de falta o exceso de inventario, optimizando los procesos de reabastecimiento y reduciendo los costos en la cadena de suministro.
- Mejora de la toma de decisiones: La exactitud en la predicción de la demanda se presenta como una herramienta esencial para la toma de decisiones fundamentadas en la gestión de la cadena de suministro, abarcando aspectos tales como: producción, planificación de la producción, adquisiciones, distribución, planificación financiera, administración de inventarios y expansión estratégica.
- Generación de ventaja competitiva: Mediante la utilización de la información generada a partir del pronóstico de la demanda, la compañía tiene la capacidad de adquirir una ventaja competitiva al alinear eficazmente sus operaciones en la cadena de suministro con las necesidades de sus clientes.
- Adaptación a los cambios del mercado: Los pronósticos efectivos de la demanda ayudan a la empresa a adaptarse rápidamente a los cambios del mercado y mantener su competitividad.

1.3 Delimitación de la Investigación

1.3.1 Delimitación espacial

La presente investigación, se realiza en la empresa CBC Peruana S.A.C, la cual se dedica a la elaboración, comercialización de bebidas PepsiCo y bebidas alcohólicas, incluyendo cerveza y bebidas carbonatadas, cuya sede administrativa se encuentra ubicada en Lima y sus plantas de producción, en Huachipa y Sullana.

1.3.2 Delimitación temporal

La recolección de la información abarca las ventas históricas de enero 2019 a julio 2023 de paquetes de gaseosa Concordia de Piña de 03 litros de 04 unidades y paquetes de bebida Evervess Giner de 1,5 litros de 06 unidades, a fin de aplicar técnicas de Machine Learning que

contribuyan a establecer un óptimo pronóstico de demanda.

1.3.3 Delimitación conceptual

El presente trabajo de investigación consiste en aplicar técnicas de Machine Learning para optimizar la planificación de la demanda de bebidas de gaseosas de la empresa CBC Peruana S.A.C, que coadyuve al manejo de operaciones, medición de capacidad de producción y la optimización de inventarios.

CAPÍTULO II: MARCO TEÓRICO

2.1 Antecedentes de la Investigación

2.1.1 Artículos

Artículo 1: Khan, P. W., Byun, Y. C., Lee, S. J., & Park, N. (2020). Machine Learning based hybrid system for imputation and efficient energy demand forecasting.

Problema de la investigación

La expansión cada vez mayor de la utilización de fuentes de energía renovable representa un desafío significativo en lo que respecta a la coordinación y gestión eficaz de la producción y distribución de energía. Es crucial perfeccionar la precisión en la previsión de la demanda energética, especialmente en el contexto de Corea del Sur, que está experimentando una transición hacia una matriz energética más sostenible. Esto, subraya la importancia de anticipar con mayor precisión los patrones de consumo de energía.

Actualmente, los métodos tradicionales de pronóstico resultan insuficientes para abordar la complejidad de los patrones de consumo, caracterizados por su naturaleza no lineal y la presencia de datos faltantes. La aplicación de técnicas convencionales de pronóstico puede carecer de la flexibilidad y eficacia necesarias para capturar las fluctuaciones en la demanda de energía en un entorno que evoluciona constantemente. La incorporación de fuentes de energía renovable intermitente, como la eólica y la solar, agrega un nivel extra de complejidad a la labor de anticipar con precisión la demanda.

La solución fundamental para enfrentar este desafío reside en la implementación de algoritmos de aprendizaje automático, una rama de la inteligencia artificial. Estos algoritmos poseen la capacidad de identificar patrones complejos y relaciones no lineales en los datos históricos de consumo de energía, lo que resulta en pronósticos más precisos y confiables. Al utilizar algoritmos de aumento de gradiente, como CatBoost y XGBoost, y aplicar técnicas de ingeniería de características, es factible superar las limitaciones de los métodos convencionales, logrando mejoras sustanciales en la calidad de los pronósticos de la demanda de energía

Objetivos

El fin principal de este artículo es exponer un modelo híbrido de predicción de la demanda de energía respaldado en tecnología de aprendizaje automático para mejorar la exactitud de la predicción y y abordar la problemática de los datos incompletos en conjuntos de datos relacionados con la energía. Para llevar a cabo este objetivo, se aplican de modo híbrido algoritmos de ascenso de gradiente (como CatBoost y XGBoost) y métodos de Random Forest

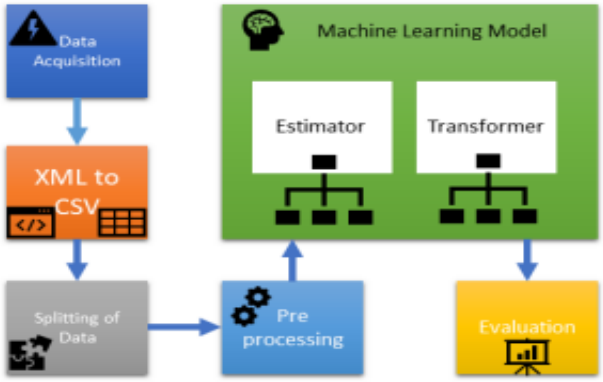
Metodología de la investigación

El enfoque propuesto combina técnicas de preprocesamiento de datos, imputación de datos faltantes y aprendizaje automático para mejorar la precisión en la predicción de la demanda de energía. Esta metodología se divide en las siguientes etapas:

a. Análisis exploratorio de datos (EDA) para comprender la distribución de los datos y detectar valores atípicos.
b. Preprocesamiento de datos, que incluye la normalización de datos y la selección de características relevantes.
c. Imputación de datos faltantes utilizando una técnica de interpolación basada en la naturaleza de los datos.
d. Entrenamiento de un modelo híbrido de aprendizaje automático que une varios algoritmos de aprendizaje automático, como arboles de decisión, redes neuronales y regresión lineal
e. Comparación del modelo propuesto con otros modelos de referencia existentes.

Figura 7

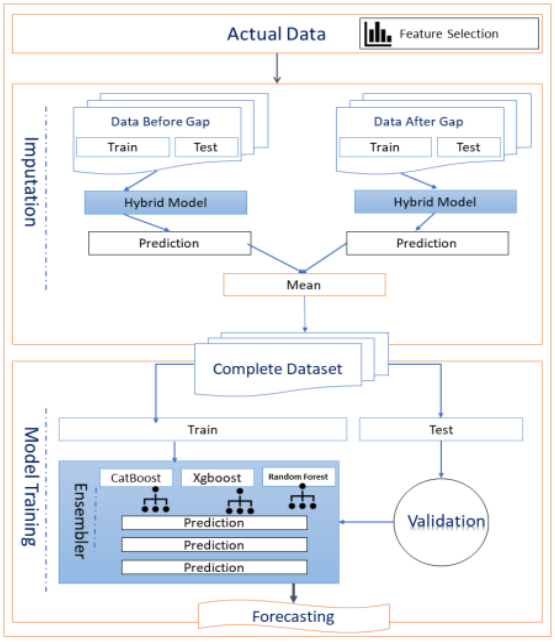
Flujo del proceso de previsión energética



Nota. Machine Learning based hybrid system for imputation and efficient energy demand forecasting (p. 10). por Khan, P, 2020.

Figura 8

Arquitectura del modelo de imputación y pronóstico propuesto



Nota. De Machine Learning based hybrid system for imputation and efficient energy demand forecasting (p. 15). por Khan, P, 2020.

Resultados y Conclusiones

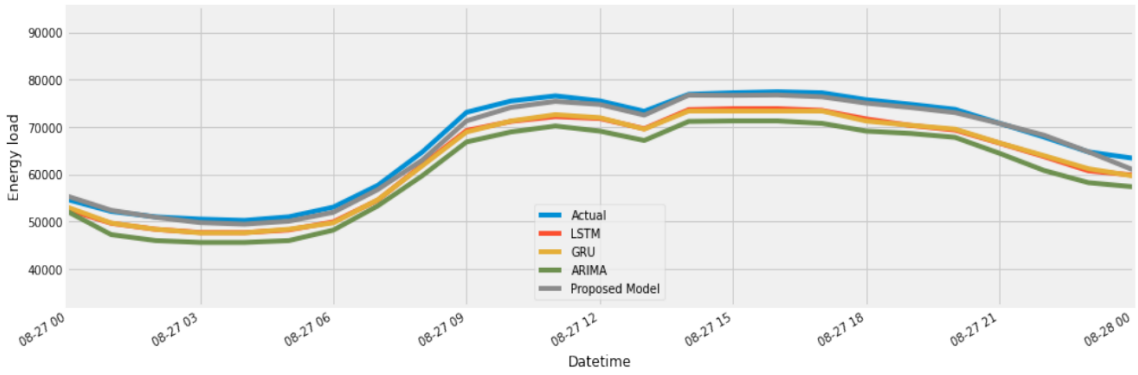
Los resultados de la investigación indican que la metodología propuesta, que utiliza técnicas de Machine Learning, es efectiva y tiene un impacto positivo en la previsión de la demanda de energía.

Algunas de las conclusiones son las siguientes:

a. Mayor Precisión en los Pronósticos: Los modelos de predicción desarrollados en base a algoritmos de aprendizaje automático como XGBoost, CatBoost y Random Forests mejoran significativamente la precisión de la predicción en comparación con los modelos tradicionales y los métodos existentes. Esto se puede ver en las métricas de evaluación: por ejemplo, el coeficiente de determinación (R-cuadrado) alcanzó un valor alto de 0,9212 en el modelo propuesto.

Figura 9

Comparación del modelo propuesto con ARIMA, GRU, LSTM



Nota. De Machine Learning based hybrid system for imputation and efficient energy demand forecasting (p. 18). por Khan, P, 2020.

b. Superación de Limitaciones: El uso de la técnica facilitó superar las restricciones propias de los procedimientos habituales en la gestión del comportamiento del consumo de energía caracterizados por su complejidad y no linealidad. Los métodos de optimización de gradiente utilizados consiguieron detectar vínculos más delicados en la información, produciendo de esta manera estimaciones más exactas y fiables.

c. Comparación con Modelos Existentes: Se compararon los resultados con los modelos y procedimientos habituales que se habían empleado para estimar el consumo de energía. Los modelos mixtos sugeridos mostraron consistentemente una mayor exactitud y eficiencia que los métodos convencionales.

Artículo 2: Miguéis, V. L (2022) Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and Machine Learning.

Problema de la Investigación

El problema central abordado en este estudio es la ineficiencia en la gestión de la demanda de pescado en el sector minorista, específicamente en el contexto de un hipermercado europeo. La ineficiencia se manifiesta en dos aspectos principales:

Existe un desperdicio significativo de pescado debido a la incapacidad de los minoristas para prever con precisión la demanda del consumidor. Esto conduce a sobreabastecimiento en ciertos días y, como resultado, el pescado no vendido se convierte en desperdicio de alimentos.

Este desperdicio no solo tiene implicaciones económicas para el minorista, sino que también contribuye a problemas medioambientales, ya que el pescado no vendido genera una huella de carbono innecesaria y desperdicia recursos naturales.

Además del problema del desperdicio, hay un desafío adicional relacionado con la censura de la demanda. La censura ocurre cuando los datos de ventas registrados no reflejan la demanda real, especialmente durante los días en que el pescado se agota en el inventario (días de censura).

La falta de datos precisos sobre la demanda real puede llevar a decisiones de reabastecimiento ineficientes, ya que los minoristas no pueden anticipar con precisión cuánto pescado se necesita realmente para satisfacer la demanda del cliente.

Objetivos de la investigación

Desarrollar modelos de pronóstico de ventas precisos y confiables para el pescado fresco, específicamente el gilthead bream (dorada), que puedan prever con exactitud la demanda de los clientes en un hipermercado europeo.

Desarrollar estrategias y enfoques para optimizar la gestión de inventarios de pescado fresco, teniendo en cuenta los patrones de demanda, las estacionalidades y las limitaciones de la cadena de suministro.

Metodología de la Investigación

a. Recopilación y Descripción de Datos

Se emplearon datos de ventas de la lubina de tamaño mediano (200-600 g) de un minorista europeo en un período de aproximadamente 2 años y 10 meses. Los datos se recopilaron desde el 1 de septiembre de 2017 hasta el 22 de octubre de 2019, excluyendo los lunes. Se realizaron análisis para comprender las fluctuaciones en las ventas, incluida la estacionalidad semanal, mensual y anual.

b. Modelado de Datos

Se generaron modelos de utilización de distintos algoritmos de aprendizaje automático, como Support Vector Regression, Random Forests, Redes Neuronales de Propagación en Capas y Long Short Term Memory, y también se aplicó el método estadístico Holt-Winters Exponential Smoothing como método de comparación. Se utilizó un enfoque de validación hacia adelante (forward validation) para prevenir el sobreajuste de los modelos.

c. Variables Predictoras

Se utilizaron variables predictoras para mejorar la precisión del pronóstico, incluyendo factores temporales (día de la semana, semana del mes, período de verano), información de precio/descuento/ventas, eventos/festividades y factores meteorológicos (precipitación y temperatura esperadas).

d. Manejo de la Demanda Censurada

Se abordó la demanda censurada (ventas que no coinciden con la demanda real debido a la falta de existencias) mediante la construcción de un modelo de pronóstico diario sin incluir datos de días con falta de existencias. Para los períodos con falta de existencias, se predijo la demanda basándose en el modelo previamente desarrollado y se seleccionó el valor más alto entre las ventas registradas y la predicción del modelo como valor de la demanda para ese día.

e. Evaluación del Rendimiento

Los modelos se evalúan utilizando medidas de precisión como el Error Cuadrático Medio (RMSE), Error Absoluto Medio (MAE), Error Positivo Medio (MPE) y Error Negativo Medio (MNE) para evaluar la precisión y los sesgos en las predicciones.

Resultados

Mejor Desempeño General: El modelo LSTM (Long Short-Term Memory) se destacó como el mejor intérprete en términos de RMSE (Root Mean Square Error) (27.82), MAE (Mean Absolute Error) (20.63), MPE (Mean Percentage Error) (17.86). Este modelo demostró una capacidad significativa para manejar datos secuenciales y predecir la demanda diaria de pescado fresco de manera precisa.

Tabla 4

Medidas de rendimiento de los modelos de predicción

Model	RMSE	MAE	MPE	MNE
Holt-Winters	40.92	29.95	37.59	24.71
RF	33.12	25.07	31.08	**20.48**
SVR	30.97	24.53	21.75	25.16
Feedforward ANN	29.03	22.25	18.04	23.89
LSTM	**27.82**	**20.63**	**17.86**	21.81
$\text{Baseline}_{\text{weekday}}$	35.96	24.01	24.57	23.52
Baseline_{t}	54.26	42.09	39.43	45.12

Nota. De Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and Machine Learning (p. 5), por Miguéis, V. L, 2022.

Comparación de Modelos Basados en Estadísticas y Aprendizaje Automático: Los modelos de aprendizaje automático superaron a los modelos basados en estadísticas como Holt-Winters. Las técnicas de aprendizaje automático, especialmente LSTM y Feedforward ANN, mostraron un mejor rendimiento en la predicción de la demanda de pescado.

Manejo de la Demanda Censurada: La consideración de la demanda censurada (días en que la demanda real no se refleja completamente en los datos de ventas debido a la falta de inventario) fue un aspecto distintivo de este estudio. Los modelos desarrollados tuvieron en cuenta esta censura, lo que mejoró la precisión de las predicciones y ayudó a reducir tanto el

desperdicio de alimentos como la falta de existencias.

Recomendación de Modelo: Dado el menor error de predicción y la capacidad para manejar tanto el exceso de oferta como la falta de oferta, se recomienda el uso del modelo LSTM para predecir la demanda de pescado fresco en este contexto específico.

Conclusiones

El estudio proporciona una metodología efectiva para el pronóstico de la demanda de pescado fresco en entornos minoristas. Utilizando técnicas de aprendizaje automático, especialmente LSTM, se alcanzó un mejor rendimiento en comparación con los modelos basados en estadísticas y los modelos de referencia. La capacidad de manejar datos secuenciales y capturar patrones complejos en los datos hizo que los modelos de aprendizaje automático fueran especialmente efectivos en este escenario.

La implementación exitosa de estos modelos tiene implicaciones significativas para la gestión de inventarios, la reducción del desperdicio de alimentos y la satisfacción del cliente. Al reducir el desperdicio y evitar la falta de existencias, las empresas minoristas pueden mejorar su rentabilidad y su impacto ambiental.

Artículo 3: Kim, S. (2023). Innovating knowledge and information for a firm-level automobile demand forecast system: A Machine Learning perspective.

Problema de la Investigación

La problemática abordada en el estudio radica en efectuar la predicción de la demanda de automóviles en la industria automotriz. La incertidumbre del comportamiento de compra de los consumidores y la necesidad de comprender los factores internos y externos que influyen en la demanda son desafíos clave enfrentados por las empresas en este sector.

Objetivos de la investigación

Evaluar y comparar varios modelos de aprendizaje automático, como Ensemble (Stack), Linear Regression, SGD, Random Forests, y Neural Networks, para determinar cuál es más efectivo en la predicción de la demanda de automóviles.

Ajustar los parámetros de los modelos de aprendizaje automático para mejorar sus capacidades predictivas. Experimentar con diferentes hiperparámetros y algoritmos para optimizar la precisión y robustez de los modelos.

Metodología de la Investigación

La metodología aplicada en el estudio se basó en la aplicación de modelos de aprendizaje automático para predecir la demanda de automóviles. A continuación, se presenta un resumen de la metodología utilizada:

Selección y Preprocesamiento de Datos: Se recopilaron datos mensuales durante un período de diez años, que incluían información sobre ventas de automóviles y varias variables endógenas (como datos mayoristas, pruebas de manejo, tráfico en la sala de exhibición) y exógenas (como precios de la vivienda, condiciones del mercado de valores, tasas de cambio, resultados de búsqueda en la web). Estos datos se preprocesaron para eliminar ruido, manejar valores atípicos y convertirlos en un formato adecuado para el análisis.

Selección de Variables: Se utilizó el método de selección de variables RReliefF para identificar las variables más relevantes que afectan la demanda de automóviles. Este método ayudó a clasificar las variables endógenas y exógenas en función de su importancia para el modelo de predicción.

Construcción de Modelos de Aprendizaje Automático: Se aplicaron cinco modelos de aprendizaje automático: Ensemble (Stack), Regresión Lineal, SGD (Gradiente Descendente Estocástico), Bosques Aleatorios y Redes Neuronales. Cada modelo se entrenó utilizando los datos preprocesados y se evaluó utilizando métricas de rendimiento como RMSE (Error Cuadrático Medio) y MAE (Error Absoluto Medio) para determinar su precisión y robustez.

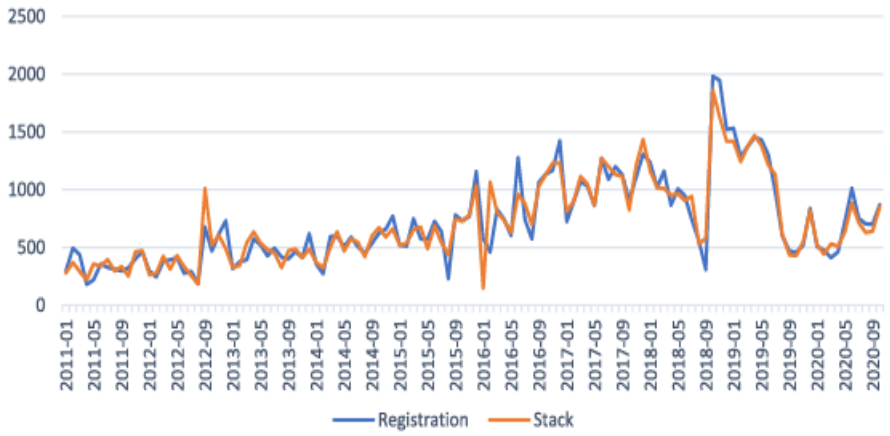
Optimización de Parámetros: Se ajustaron los parámetros de los modelos para mejorar su precisión. Se realizaron experimentos para encontrar la combinación óptima de hiperparámetros que proporcionara los mejores resultados predictivos.

Evaluación del Modelo: Se compararon las métricas de rendimiento de los modelos para identificar el modelo más confiable en términos de precisión y capacidad predictiva. Se realizaron análisis de sensibilidad para evaluar cómo los cambios en variables clave afectaban las predicciones de demanda.

Análisis e Interpretación de Resultados: Se analizaron los resultados para identificar patrones, tendencias y relaciones entre las variables. Se interpretaron los hallazgos para comprender mejor los factores que afectan la demanda de automóviles y se proporcionaron recomendaciones basadas en los resultados del estudio.

Figura 10

Resultado de la predicción- Stack



Nota. De Innovating knowledge and information for a firm-level automobile demand forecast system: A Machine Learning perspective (p. 10) por Kim, S, 2023.

Resultados y Conclusiones

Se evaluaron cinco modelos de aprendizaje automático. El modelo de Ensemble (Stack) mostró el menor RMSE, con un valor de 0.424, indicando la menor discrepancia entre las predicciones y los valores reales de las ventas de automóviles. Le siguieron la Regresión Lineal (RMSE: 0.576), Random Forests (RMSE: 0.774), SGD (RMSE: 0.827) y Redes Neuronales (RMSE: 0.904).

Al analizar el MAE, que mide el promedio de las discrepancias absolutas entre las predicciones y los valores reales, el modelo de Ensemble (Stack) demostró una vez más el menor error (0.173). Le siguieron Regresión Lineal (MAE: 0.210), Bosques Aleatorios (MAE: 0.674), SGD (MAE: 0.790) y Redes Neuronales (MAE: 0.980).

Tabla 5

Comparación de modelos por RMSE

	Stack	Linear regression	SGD	Random forest	Neural network
Stack	-	0.424	0.173	0.226	0.026
Linear Regression	0.576	-	0.210	0.326	0.020
SGD	0.827	0.790	-	0.508	0.057
Random Forest	0.774	0.674	0.492	-	0.147
Neural Network	0.904	0.980	0.943	0.853	-

* Table shows probabilities that the score for the model in the row is higher than that of the model in the column

Nota. De Innovating knowledge and information for a firm-level automobile demand forecast system: A Machine Learning perspective (p. 25) por Kim, S, 2023.

Un enfoque de modelo conjunto (Ensemble) y la inclusión cuidadosa de variables endógenas y exógenas son cruciales para predecir con precisión la demanda de automóviles. La Regresión Lineal, a pesar de su simplicidad, demostró ser un modelo eficaz en este contexto específico.

2.1.2 Tesis relacionadas

Tesis 1: Lara S (2022). "Aplicación de Técnicas de Machine Learning como método de validación para predecir la efectividad de un modelo estadístico de series de tiempo en la producción de fruta fresca en las diferentes provincias del Ecuador"

Problema principal

Los agricultores de frutas frescas en Ecuador enfrentan desafíos significativos debido a la falta de acceso a información precisa y tecnológica. Esta limitación impide que se tomen decisiones informadas sobre la producción de cultivos, lo que se vuelve aún más crucial en un contexto de cambio climático y variabilidad en las condiciones agrícolas. La falta de educación formal, la escasa automatización en los procesos agrícolas y la ausencia de modelos predictivos actuales crean obstáculos para la optimización de la producción y la planificación eficiente.

Objetivos de la Investigación

- Elaborar un enfoque predictivo para evaluar la eficacia de modelos de series temporales en la producción de frutas frescas en Ecuador, mediante la aplicación de técnicas y

modelos de aprendizaje automático supervisado.

- Recabar datos bibliográficos acerca de las variables que impactan en la producción agrícola de frutas frescas, así como explorar herramientas basadas en algoritmos de aprendizaje automático con el propósito de anticipar la producción de frutas frescas en las provincias de Ecuador.

- Obtener una base de datos del Sistema de Información Pública y Agropecuaria del Ministerio de Agricultura y Ganadería, focalizando la atención en información correspondiente a 08 cosechas de frutas frescas, y analizar las variables más relevantes con el fin de construir algoritmos predictivos específicamente elegidos.

- Emplear modelos estadísticos de series temporales como base para los algoritmos de aprendizaje automático, con el propósito de anticipar la producción de frutas frescas en Ecuador.

Metodología

Selección de Datos: Para llevar a cabo la investigación, se seleccionó un conjunto de datos históricos confiables y relevantes sobre la producción de banano en Ecuador. Estos datos pueden provenir de fuentes gubernamentales, organizaciones agrícolas o instituciones de investigación. Es crucial que los datos sean completos, precisos y actualizados para garantizar la calidad de los resultados.

Preprocesamiento de Datos: Antes de utilizar los datos para el análisis, se realiza un proceso de preprocesamiento para limpiar y preparar los datos. Esto implica identificar y manejar valores atípicos, gestionar datos faltantes o nulos, normalizar las variables si es necesario y convertir datos categóricos en una forma adecuada para el análisis. El objetivo es tener un conjunto de datos limpio y homogéneo para entrenar los modelos.

Construcción del Modelo: En esta etapa, se elige el tipo de modelo de aprendizaje automático que se utilizará para predecir la producción de banano. En este caso, se opta por las redes neuronales feedforward con embeddings. Se configuran los hiperparámetros del modelo, como el número de capas, el número de neuronas en cada capa, la función de activación y el optimizador. Además, se decide la estructura del modelo, incluyendo las entradas (variables independientes) y la salida (producción de banano).

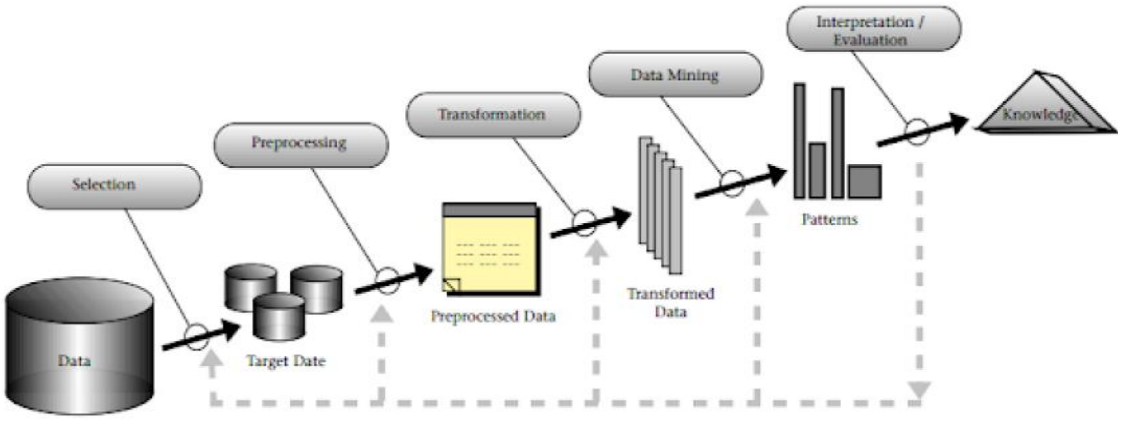
Validación y Evaluación: Una vez que el modelo se entrena con los datos, se procede a su validación y evaluación. Si los resultados no son satisfactorios, se ajustan los hiperparámetros o se consideran otras técnicas para mejorar la precisión.

Implementación del Modelo: Después de validar y evaluar el modelo, se implementa para realizar predicciones futuras de la producción de banano en Ecuador. El modelo entrenado se utiliza para predecir la producción en un período futuro específico, basándose en las variables de entrada proporcionadas. Las predicciones se pueden utilizar para la planificación agrícola, la toma de decisiones empresariales y otras aplicaciones prácticas.

Documentación y Presentación: Es fundamental documentar todo el proceso, incluyendo los detalles del modelo, los resultados obtenidos, las métricas de evaluación y cualquier otra información relevante. Además, se prepara una presentación clara y comprensible para comunicar los resultados a audiencias no técnicas, lo que puede incluir gráficos, visualizaciones y explicaciones concisas del proceso y los hallazgos.

Figura 11

Componentes del proceso KDD



Nota. De Aplicación de Técnicas de Machine Learning como método de validación para predecir la efectividad de un modelo estadístico de series de tiempo en la producción de fruta fresca en las diferentes provincias del Ecuador (p.12), por Lara, S, 2022.

Resultados y Conclusiones

El modelo que utiliza embeddings logra una mejora significativa, ya que el validation_loss disminuye de 0.37 a 0.27 y el validation_MSE de 0.17 a 0.09.

Tabla 6

Resultados de los modelos de predicción

Redes FeedForward	loss	val_loss	mse	val_mse
Multivariable	0.4991	0.3700	0.3259	0.1738
Embedding	0.4128	0.2755	0.2175	0.0947

Nota. De Aplicación de Técnicas de Machine Learning como método de validación para predecir la efectividad de un modelo estadístico de series de tiempo en la producción de fruta fresca en las diferentes provincias del Ecuador (p.11), por Lara, S, 2022.

Se realizaron análisis estadísticos, incluido el análisis de correlación de Pearson, para seleccionar variables pertinentes y entender las relaciones entre ellas. Estos análisis proporcionaron información valiosa sobre la producción agrícola y ayudaron en la elección de las variables para los modelos de predicción.

Se demostró la factibilidad de elaborar redes neuronales artificiales del tipo MLP para predecir la producción agrícola. Específicamente, los modelos de redes neuronales MLP, incluyendo Feedforward Multivariate y Feedforward Embeddings, demostraron una capacidad sólida para predecir las series de tiempo de la producción de cultivos de frutas frescas en Ecuador. Entre estos, el modelo Feedforward Embeddings demostró una mayor precisión en las predicciones.

Tesis 2: Bastarrica Lacalle, D. F. (2020). Predicción de series temporales mediante el método k-NN: explicabilidad y algoritmos de ensamblado.

Problema de la investigación

En el contexto actual de creciente digitalización, la abundancia de datos disponibles abre nuevas posibilidades para la toma de decisiones informadas y la anticipación de eventos futuros. La predicción de series temporales, que implica el análisis y la proyección de datos a lo largo del tiempo, emerge como una herramienta esencial en diversos campos, desde la planificación

meteorológica hasta la gestión de recursos esenciales y la anticipación de patrones de enfermedades.

No obstante, a pesar de los avances en técnicas de predicción, persisten desafíos significativos. En este contexto, la presente investigación se propone abordar una problemática central relacionada con la predicción de series temporales, específicamente mediante el uso del método de los k-vecinos más cercanos (k-NN). Aunque el k-NN se ha destacado por su simplicidad y rendimiento, se plantea la cuestión de su capacidad para proporcionar predicciones robustas y transparentes, así como para adaptarse a diversos comportamientos de las series temporales

Objetivos de la investigación

- Implementar una estrategia de boosting para k-NN en series temporales.
- Mejorar la eficiencia de la estrategia de boosting implementada.
- Incrementar la comprensibilidad de las predicciones generadas por k-NN, incorporando métricas particulares y visualizando los k-vecinos seleccionados.
- Ampliar la funcionalidad de la aplicación web existente para incluir la predicción mediante la combinación de diferentes métodos, integrando visualmente las componentes explicativas.
- Realizar un rediseño completo de la aplicación web, enfocándose en perfeccionar la experiencia de usuario y acortar los tiempos de respuesta.

Metodología

Cada serie temporal se dividió proporcionalmente en warm-up (60%), en datos de entrenamiento (30%), y datos de prueba (10%). Se utilizó la misma configuración para todas las series, incluyendo valores de 1 a 50 para la búsqueda de parámetros, la ponderación de pesos proporcional a la distancia para el cálculo de la predicción y el error RMSE para medir sobre la predicción. Asimismo, se definieron manualmente semillas para la generación de valores aleatorios, asegurando replicabilidad en los resultados. Luego, se utilizaron métodos de referencia como Naive, Seasonal Naive, y k-NN con configuración estándar para comparar el rendimiento de los Random Neighborhoods.

Resultados y Conclusiones

Para los errores de Predicción (RMSE), los métodos de Random Neighborhoods superan a los métodos de referencia (Naive y S.Naive) en las series temporales, con mejoras significativas, especialmente en el método k-NN básico. El método de Bloques destaca con el mejor rendimiento tanto en los conjuntos de entrenamiento como en los de prueba.

Tabla 7

Estadístico comparativo DM para la serie Aire

Método	Entrenamiento			Test		
	Naive	S. Naive	k-NN	Naive	S. Naive	k-NN
Bloques	9,9928	25,438	7,6305	6,7004	21,172	3,6816
Muestras	9,5668	25,435	6,3429	6,275	21,194	3,05155
Fitness	10,13	25,587	7,2129	6,4016	21,317	2,9733
Mínimos	9,2673	25,633	6,5939	6,4745	21,41	3,5167

Nota. Bastarrica Lacalle, D. F. (2020). Predicción de series temporales mediante el método k-NN: explicabilidad y algoritmos de ensamblado

El método de *Fitness* muestra algunas limitaciones para superar a otros métodos, aunque supera al k-NN base en ciertos casos.

Los métodos de Random Neighborhoods se destacan en comparación con los métodos de referencia, y cada uno tiene sus fortalezas según la serie temporal. Aunque el método de *Fitness* tiene algunas limitaciones, los resultados indican que estos enfoques son prometedores para la predicción de series temporales.

Tesis 3. Correa A. (2023). Análisis de modelos basados en Machine Learning para la predicción de la demanda de productos en la empresa Dyna & Cía. S.A

Problema de la investigación

El problema principal radica en la imprevisibilidad de la demanda de los productos. La demanda no sigue un patrón lineal y está influenciada por diversos factores, como aumentos inesperados en el consumo, variaciones climáticas, eventos especiales, y cambios económicos, entre otros. Esta variabilidad hace que sea difícil prever con precisión cuánto de cada producto se venderá en un período determinado.

La empresa ha intentado abordar este problema aumentando los niveles de inventario, lo que implica una inversión significativa de capital. Sin embargo, esta estrategia no ha sido completamente efectiva, ya que los agotados de productos todavía ocurren, lo que resulta en pérdida de ventas y clientes insatisfechos.

Objetivo de la investigación

- Analizar el comportamiento del uso de técnicas y modelos basados en Machine Learning para el pronóstico de la demanda de productos en la empresa Dyna & Cía. S.A.
- Examinar el modelo tradicional actualmente utilizado por la empresa para la predicción de la demanda.
- Identificar modelos de Machine Learning útiles para el trabajo con series de tiempo y generación de pronósticos.
- Evaluar el rendimiento de modelos Machine Learning en términos de reducción de la diferencia entre lo pronosticado y lo real.
- Comparar los resultados y sintetizar conclusiones obtenidas, proporcionando recomendaciones prácticas y consideraciones para la generación de pronósticos de la demanda

Metodología de la investigación

El estudio se basa en el enfoque de Machine Learning, específicamente en técnicas de Aprendizaje Automático. A continuación, se detallan los pasos seguidos:

a. Adquisición de Datos:

Se identificaron fuentes relevantes de información, incluyendo históricos de pedidos, maestro de productos y unidades de empaque. La información fue extraída del ERP de la compañía y se almacenó en archivos CSV.

b. Procesado de Datos:

Se utilizó Python y la biblioteca Pandas para consolidar y procesar los datos. Esto incluyó la transformación de variables, manejo de datos nulos y la unión de características relevantes,

preparando los datos para el análisis y entrenamiento de modelos de Machine Learning.

c. Extracción y Creación de Variables Significativas:

Se seleccionaron variables relevantes y se crearon nuevos campos según el contexto del estudio. Se dividió el conjunto de datos en grupos de entrenamiento y prueba, conservando el orden cronológico para las series temporales.

d. Algoritmo de Machine Learning:

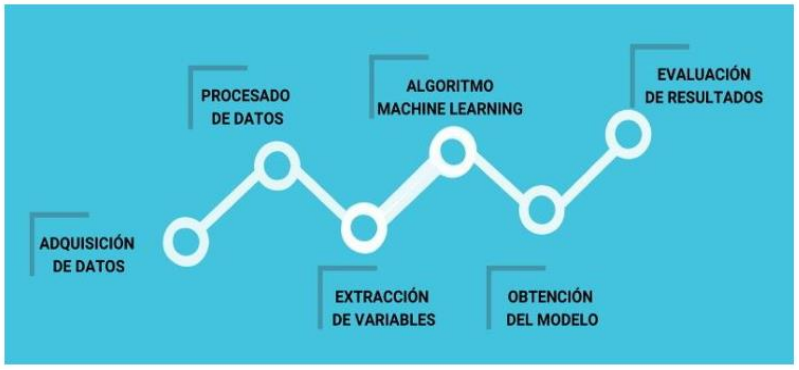
Se exploraron varios modelos de Machine Learning: ARIMA, Redes Neuronales, Regresión y Random Forest. Cada modelo fue seleccionado por sus características y adaptabilidad a los datos. Los modelos se implementaron utilizando las librerías Scikit-Learn y Tensorflow.

e. Evaluación y Obtención de Modelos:

La métrica MAE (Error Medio Absoluto) se utilizó para evaluar la precisión de los modelos. Se realizaron iteraciones para optimizar los hiperparámetros y se utilizaron herramientas como GridSearch para mejorar la precisión de las predicciones. Las métricas obtenidas de cada modelo fueron comparadas para determinar su eficacia en la predicción de la demanda de productos.

Figura 12

Metodología Machine Learning



Nota. De análisis de modelos basados en Machine Learning para la predicción de la demanda de productos en la empresa Dyna & Cía. S.A(p.31), por Correa, A, 2023.

Resultados y Conclusiones

A continuación, el detalle de los resultados y conclusiones:

- Para el modelo ARIMA se obtuvo un MAE de 146 unidades, mostrando la mejor precisión en la predicción de la demanda para este producto.
- El modelo de Redes Neuronales registra un MAE de 6, 230 unidades, presentando un buen ajuste a los datos reales, evitando suavizaciones excesivas y capturando mejor los picos de demanda.
- La Red Neuronal también mostró el mejor rendimiento para este producto, con un MAE de 11, 175 unidades, aproximadamente un 15% mejor que el método actual.
- Todos los modelos de Machine Learning superaron significativamente el método actual de la empresa, reduciendo el MAE en promedio en un 23%. Esto demuestra la eficacia de los modelos de Machine Learning en comparación con los métodos tradicionales de pronóstico.
- Los modelos de Machine Learning, especialmente las Redes Neuronales, se destacaron en la predicción de la demanda. Tienen la capacidad de aprender patrones complejos y adaptarse a las variaciones en la demanda, proporcionando pronósticos más precisos y ajustados a la realidad.
- La implementación de modelos de Machine Learning tiene un impacto directo en la gestión del inventario. Los pronósticos más precisos permiten una planificación más efectiva, reduciendo la pérdida de ventas por falta de disponibilidad de productos.
- El uso del MAE como métrica de evaluación simplifica la comprensión de la precisión del modelo. Esta métrica se traduce directamente en unidades, facilitando la comunicación de los resultados a personas no familiarizadas con conceptos de Machine Learning.

2.2 Bases Teóricas

En esta sección se presentarán los conceptos clave que se utilizarán en este trabajo de investigación en los capítulos siguientes, y se proporcionarán definiciones de estos conceptos basadas en las perspectivas de varios autores para obtener una comprensión completa de los mismos.

2.2.1. Inteligencia Artificial

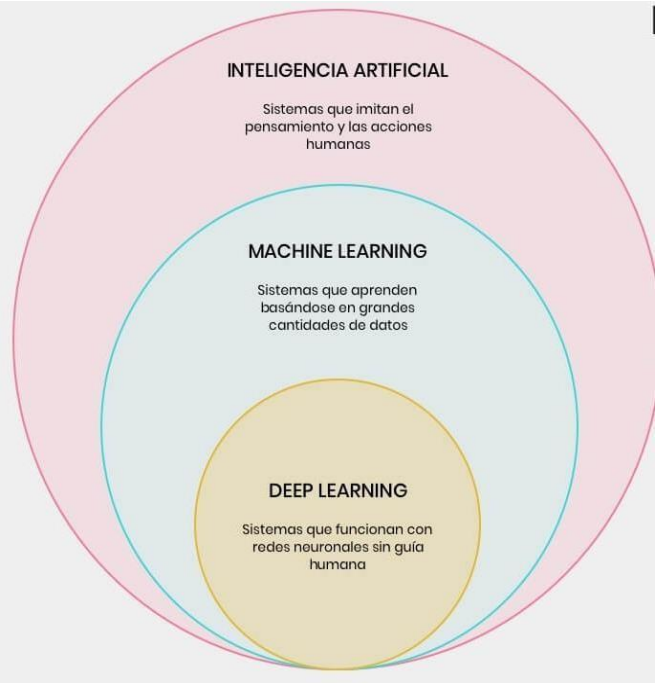
El concepto de inteligencia artificial (IA) es una disciplina más de la informática cuyo enfoque es la creación de sistemas que pueden llegar a imitar la inteligencia humana. Según John McCarthy, uno de los fundadores de la IA, la define como "la ciencia y la ingeniería de la creación de máquinas inteligentes, especialmente programas informáticos inteligentes". (McCarthy, J, 2007).

Asimismo, Stuart Russell y Peter Norvig lo definen como "el estudio de cómo hacer que las computadoras hagan cosas que los humanos actualmente hacen mejor". Por otro lado, Tom Mitchell lo define como "el estudio de algoritmos que permiten a las computadoras aprender a partir de datos". (Russell, S. J, 2010)

La inteligencia artificial se está convirtiendo en un instrumento primordial para una variedad de campos, como la ingeniería, la robótica y la medicina. Sin embargo, la medida también genera inquietudes sobre su posible impacto en el empleo y la privacidad, según se ha señalado.

Figura 13

Inteligencia artificial



Nota. Adaptado de inteligencia artificial, por López, S, 2020.

2.2.2 Machine Learning

"El Machine Learning consiste en la creación de modelos o algoritmos para analizar datos, aprender de ellos y, luego, hacer una predicción de su posible comportamiento en un rango de tiempo o situación estimada". (Martínez, W. R., 2020)

El ML se ha transformado en una herramienta primordial para muchos campos. "En la actualidad, encontramos varias aplicaciones de la Inteligencia artificial, a través del Machine Learning, en la ciberseguridad informática, entre ellas: detección de fraude de tarjetas bancarias, detección de intrusos, clasificación de malware y detección de ataques". (Martínez, W. R, 2020)

"La aplicación de estas técnicas tiene un alto potencial de uso en la producción agropecuaria, ya que posibilita el desarrollo de sistemas inteligentes de apoyo a las decisiones productivas". (Ramírez Morales, I, 2018)

"Los modelos de fiabilidad avanzados y algoritmos de predicción de fallos pueden facilitar a los operadores la detección anticipada de fallos de componentes en los aerogeneradores y, en base a ello, adaptar sus estrategias de mantenimiento". (Reder, M. D, 2018)

2.2.3 Deep learning

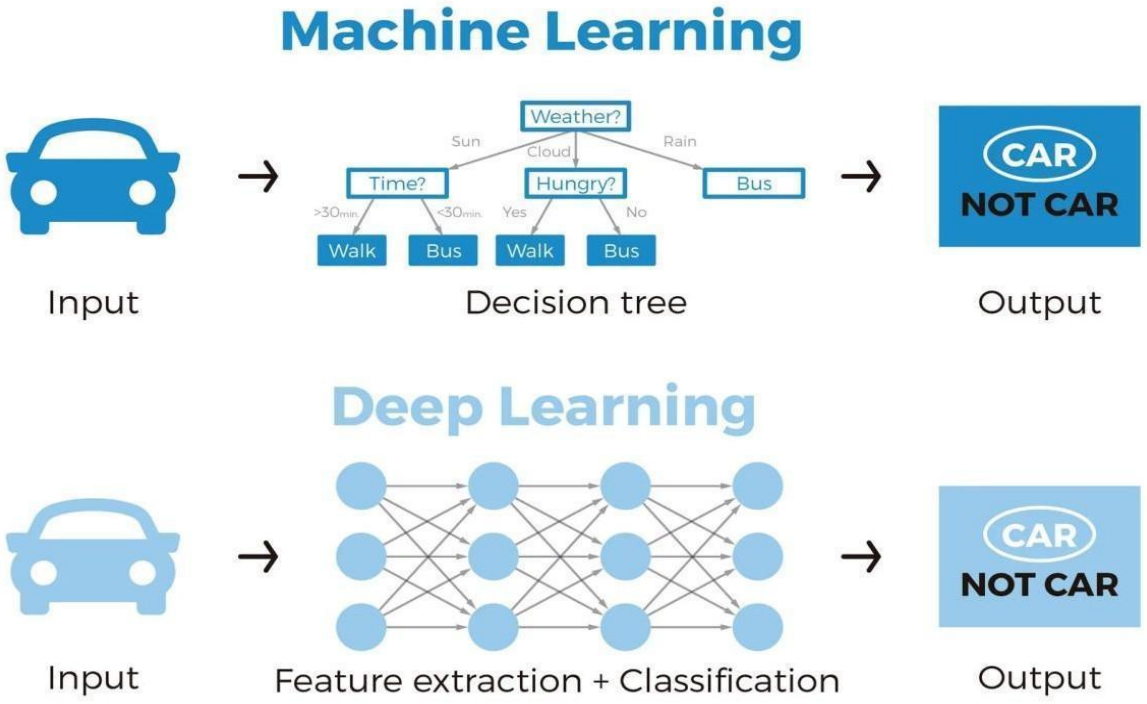
"Deep Learning es una técnica de Machine Learning que utiliza redes neuronales artificiales con múltiples capas para aprender y realizar tareas complejas, como reconocimiento de voz, visión por computadora y procesamiento del lenguaje natural". (Martínez, W. R. A, 2020)

"Las redes neuronales artificiales son un conjunto de algoritmos y técnicas que se utilizan para modelar patrones complejos y relaciones en los datos". (Ramírez Morales, I, 2018)

"El Deep Learning ha permitido avances significativos en la visión por computadora, como la detección de objetos y la segmentación de imágenes. La mejora en la exactitud del procesamiento del lenguaje natural ha resultado en el desarrollo de asistentes virtuales y chatbots más eficientes”. (Reder, M. D, 2018)

Figura 14

Diferencias entre Machine Learning y Deep Learning



Nota. Adaptado de Machine Learning, por Bismart, 2022

2.2.4 Tipos de aprendizaje y algoritmos

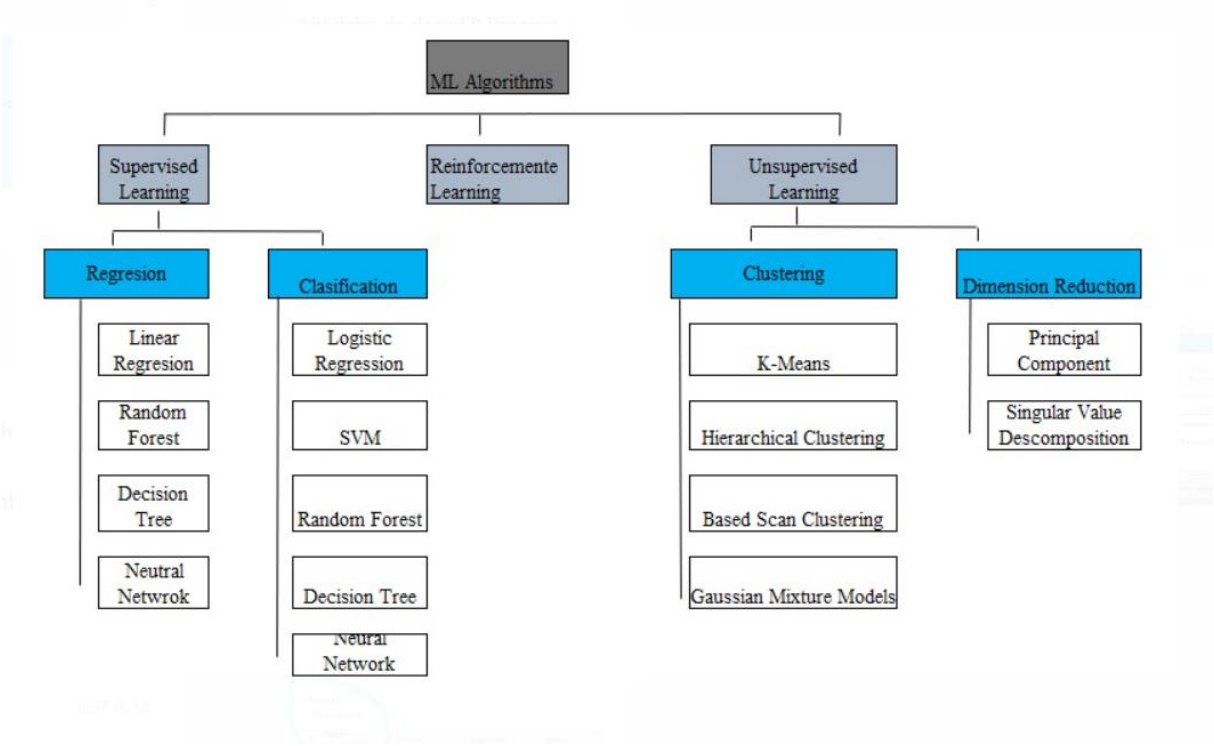
Para Benos et al. (2021), El aprendizaje automático, que se encuentra dentro de la subárea de Machine Learning, comúnmente se organiza en distintas categorías, las cuales son:

- Aprendizaje Supervisado: “Es una técnica de aprendizaje automático en la que se entrena un modelo utilizando un conjunto de datos etiquetados. El objetivo del aprendizaje supervisado es que el modelo pueda generalizar y producir salidas necesarias para nuevas entradas que no se han visto antes.” (Benos, L.,2021).
- Aprendizaje No supervisado: “Es una forma de aprendizaje automático en la que se utilizan algoritmos para analizar datos sin necesidad de codificación o clasificación. El objetivo del aprendizaje no supervisado es encontrar patrones y tendencias ocultos en los datos que puedan ayudar a tomar decisiones o identificar patrones.” (Benos, L.,2021).

- Aprendizaje Reforzado: “Es una clase de aprendizaje automático en el que un agente aprende a tomar decisiones en un entorno determinado a través de la interacción continua con ese entorno.” (Benos, L.,2021).

Figura 15

Tipos de Aprendizaje algoritmos Machine Learning



Nota. Adaptado de Tipos de Aprendizaje algoritmos Machine Learning, por Pugliese, L, 2021.

2.2.5 Forecasting y Regresión

“Forecasting es una técnica de pronóstico de demanda que se basa en el análisis de datos históricos y en la identificación de patrones y tendencias en esos datos, con el objetivo de predecir futuras condiciones de demanda y ayudar a tomar decisiones empresariales informadas”. (Petropoulos et al., 2022).

“La Regresión es un método estadístico empleado para examinar la relación lineal entre dos o más variables, centrándose en describir y anticipar la conexión entre estas variables en futuras instancias”. (Petropoulos et al., 2022).

La diferencia entre estos dos conceptos en el ámbito del aprendizaje automático radica en sus propósitos específicos. “El Forecasting es una técnica que se utiliza para predecir los valores futuros de una serie de tiempo basándose en un número limitado de puntos de datos históricos, mientras que la regresión es un método de aprendizaje supervisado que se utiliza para modelar las relaciones entre las variables de entrada y salida”. (Petropoulos et al., 2022).

2.2.6 Regresión Lineal

“La regresión lineal constituye un modelo estadístico destinado a identificar la relación lineal entre una variable dependiente (o de respuesta) y una o más variables independientes (o predictoras). Tiene como objetivo, encontrar la línea recta que mejor se ajusta a los datos, lo que permite predecir la variable dependiente en función de las variables independientes y comprender la naturaleza de la relación entre ellas.” (Mendenhall, W., Beaver, R. J., & Beaver, B. M., 2017).

Fórmula 1. Regresión Lineal

$$Y = b_0 + b_1 X$$

Donde:
Y representa la variable dependiente predicha
b_0 es el término de intercepción
b_1 es el coeficiente de pendiente
X es la variable independiente.

2.2.7 Autoregresión

Según Shumway, R. H., & Stoffer, D. S, (2010), señalan que, la autoregresión (AR) es un método estadístico utilizado en el análisis de series temporales que modela la relación entre una variable en un período de tiempo específico y sus valores anteriores en el tiempo. En un modelo AR, se supone que el valor actual de la variable depende linealmente de sus valores previos, con coeficientes específicos para cada retraso temporal. En esencia, un proceso autorregresivo se basa en su propio historial, donde el valor actual es una combinación lineal de sus valores pasados.

Fórmula 2. Autoregresión

$$X_t = \emptyset_1 X_{t-1} + \varepsilon_t$$

Donde:

X_t es el valor en el tiempo *t*

$\emptyset_1$ es el coeficiente de autoregresión de primer orden

X_{t-1} es el valor en el tiempo $t-1$

ε_t es el término de error en el tiempo t

2.2.8 Series Temporales

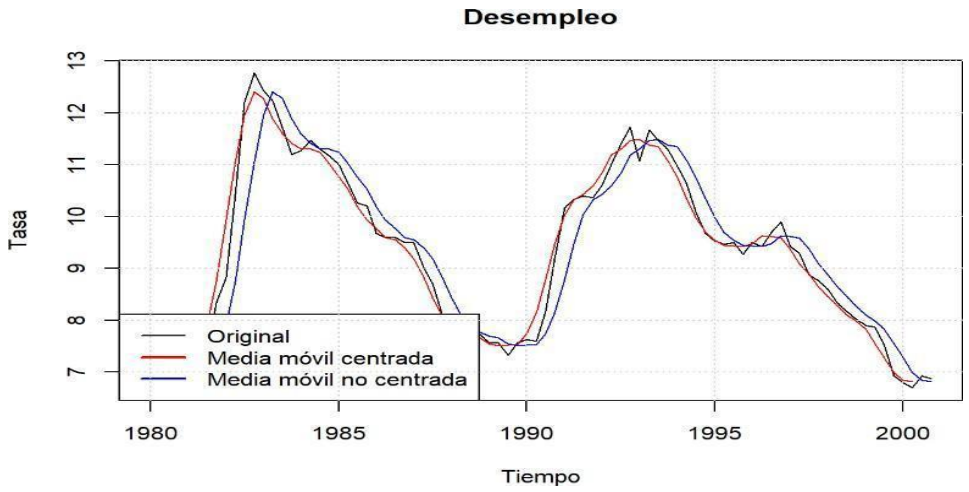
Según Parra, F. (2019), "Una serie de tiempo es una serie de observaciones de una variable realizadas en intervalos regulares."

"Las series de tiempo se caracterizan por la dependencia del tiempo, lo que significa que los valores presentes están relacionados con los valores pasados y pueden afectar los valores futuros." (Martínez, W. R, 2020)

Las series temporales pueden variar según el intervalo de tiempo utilizado para recoger los datos, como días, semanas, meses o años. Estas se caracterizan por que pueden presentar componentes de Tendencia, Estacionalidad y Ciclos.

Figura 16

Representación gráfica de la variable



Nota. Adaptado de Econometría Aplicada II (p. 40) por Parra, F. (2019)

2.2.8.1 Métodos ARIMA y SARIMA

Los métodos más comunes en el uso de la técnica de series de tiempo y la predicción incluyen ARIMA y SARIMA, que se describen a continuación.

Modelo ARMA

El modelo ARMA (media móvil autorregresivo) se trata de un modelo estadístico empleado para analizar y prever patrones en series temporales. "El modelo ARMA se utiliza para modelar la dependencia temporal en series temporales y puede adaptarse a diferentes tipos de datos de series temporales. Este se representa como ARMA *(p, q*), donde p es el orden del modelo *AR* y *q* es el orden del modelo MA." (Shumway, R. H., & Stoffer, D. S., 2010).

Fórmula 3. Fórmula ARMA

$$X_t = c + \phi_1 X_{t-1} + \phi_1 X_{t-2} + \cdots + \phi_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Donde:

X_t es el valor en el tiempo t
p es el orden del componente autorregresivo
q es el orden del componente de media móvil
c es una constante
ϕ_1 son los coeficientes autorregresivos
θ_j son los coeficientes de la media móvil
ε_t es el término de error en el tiempo t

Modelo ARIMA

"Es un modelo estadístico utilizado para el análisis y predicción de series temporales. Este modelo utiliza información histórica de una serie temporal para poder predecir sus futuras acciones." (Shumway, R. H., & Stoffer, D. S, 2006)

El modelo está comprendido por 3 componentes:

a. (AR) componente autorregresivo

b. (MA) componente de media móvil

c. (I) componente diferencial

Fórmula 4. Fórmula ARIMA

$$\nabla_d X_t = c + \phi_1 \nabla_d X_{t-1} + \phi_2 \nabla_d X_{t-2} + \cdots + \phi_p \nabla_d X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_2 \varepsilon_{t-q} + \varepsilon_t 2$$

Donde:
d es el orden de diferenciación, lo que significa que la serie original Xt se diferencia "*d*" veces.
∇ representa el operador de diferencia, que se utiliza para calcular las diferencias entre los valores de la serie temporal.
c es una constante
$\phi_1 \ldots \phi_p$ son los coeficientes autorregresivos
$\theta_j \ldots \theta_p$ son los coeficientes de la media móvil
ε_t es el término de error en el tiempo *t*

El componente diferencial, "*d*", indica cuántas veces debes diferenciar la serie temporal para lograr la estacionariedad. En otras palabras, "*d*" representa el número de diferencias que se toman para convertir la serie en una serie estacionaria. Esto implica restar cada valor de la serie temporal del valor anterior.

Modelo SARIMA

SARIMA (Promedio móvil integrado autorregresivo estacional) es una extensión del modelo ARIMA diseñada para la modelación y predicción de series temporales que muestran comportamientos estacionales. Según Smith, J. (2005), el modelo SARIMA es especialmente útil cuando los datos muestran estacionalidad, es decir, patrones repetitivos a lo largo del tiempo, como las ventas mensuales de productos, las temperaturas estacionales, los datos financieros, entre otros. (Smith, J.2017, p. 87-102)

Fórmula 5. Fórmula SARIMA

$$\nabla_d \nabla_s DX_t = c + \phi_1 \nabla_d \nabla_s DX_{t-1} + \phi_2 \nabla_d DX_{t-2} + \cdots + \phi_p \nabla_d X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_2 \varepsilon_{t-q} + \varepsilon_t$$

Donde:

∇d *y* ∇s son operadores de diferencia que se utilizan para calcular las diferencias de la serie temporal con respecto al tiempo y a la estacionalidad, respectivamente.

d es el orden de diferenciación no estacional

D es el orden de diferenciación estacional

p es el orden del componente autorregresivo

q es el orden del componente de media móvil

c es una constante

ϕi son los coeficientes autorregresivos

θj son los coeficientes de la media móvil

εt es el término de error en el tiempo *t*.

2.2.8.2 FB Prophet

FB Prophet es un procedimiento para pronosticar datos de series de tiempo creado por Core Data Science de Facebook. Su objetivo es poder realizar pronósticos "a escala", lo que significa que FB Prophet quiere ser la herramienta de pronóstico automatizada en la naturaleza, lo que brinda más facilidad de uso para ajustar los métodos de series de tiempo y permite a los analistas de cualquier experiencia o personas con poco o (posiblemente) ningún conocimiento previo en pronósticos para poder realizar pronósticos con éxito. (Christophorus Beneditto, A., Darmawan, W., Bellatasya Unrica, N., Novita, H., 2021)

Según Facebook, FB Prophet “funciona mejor con series temporales que tienen fuertes efectos estacionales y varias temporadas de datos históricos y es resistente a valores atípicos y cambios en la tendencia”. Su naturaleza automática da flexibilidad a los datos de series temporales que tienen cambios dramáticos y, por lo tanto, los analistas no tienen que preocuparse de que sus datos no sean adecuados para pronóstico con FB Prophet.

A continuación, algunas características clave de Prophet, según Christophorus Beneditto,

A., Darmawan, W., Bellatasya Unrica, N., Novita, H. (2021):

a. Manejo de la Estacionalidad: Prophet puede manejar tanto la estacionalidad anual como la semanal, lo que lo hace adecuado para datos con múltiples patrones estacionales. También puede manejar días festivos y eventos especiales.
b. Modelado de Tendencias: Prophet utiliza una curva de crecimiento lineal o logístico segmentada para modelar la tendencia subyacente en los datos. También puede capturar automáticamente los puntos de cambio de tendencia.
c. Efectos de Días Festivos: Puede especificar días festivos y eventos que podrían afectar sus datos de series temporales, y Prophet incorporará estos efectos en sus pronósticos.
d. Estimación de Incertidumbre: Prophet proporciona intervalos de incertidumbre alrededor de sus pronósticos, lo que ayuda a los usuarios a evaluar la confiabilidad de las predicciones.
e. Personalización: Los usuarios pueden personalizar varios hiper parámetros del modelo para ajustar el proceso de pronóstico.
f. Facilidad de Uso: Prophet está diseñado para ser fácil de usar y no requiere que los usuarios tengan un amplio conocimiento en pronóstico de series temporales. Es particularmente útil para analistas de negocios y científicos de datos que desean generar rápidamente pronósticos.

2.2.9 Criterio de información de Akaike (AIC)

El criterio de información de Akaike (AIC) es una métrica utilizada para comparar y seleccionar modelos estadísticos. Fue desarrollado por el estadístico japonés Hirotugu Akaike. “El AIC combina la capacidad de ajuste del modelo con su complejidad, lo que lo convierte en una herramienta importante para seleccionar el modelo más apropiado entre varios candidatos.” (Burnham, K. P., & Anderson, D. R, 2002)

Fórmula 6. Fórmula SARIMA

$$AIC = -2\ln(L) + 2k$$

Donde:

AIC es el Criterio de Información de Akaike.
L representa la función de verosimilitud del modelo, que mide cuán bien el modelo se ajusta a los datos observados.
k es el número de parámetros libres en el modelo.

El objetivo del AIC, consiste en encontrar un equilibrio entre el ajuste del modelo a los datos (medido por la función de verosimilitud) y la penalización por la complejidad del modelo (a través del término 2k). El modelo que presenta el menor valor de AIC se considera el más adecuado en cuanto a ajuste y complejidad Adicionalmente, señalan que, AIC se utiliza ampliamente en estadísticas, econometría y otras disciplinas para seleccionar modelos, como modelos de regresión, series temporales, y más, ayudando a los analistas a elegir modelos que expliquen de manera efectiva los datos sin ser demasiado complejos. (Burnham, K. P., & Anderson, D. R, 2002)

2.2.10 Dickey-Fuller Aumentada (ADF)

"La prueba de Dickey-Fuller aumentada (ADF) es una herramienta estadística fundamental en el análisis de series temporales utilizada para evaluar la estacionariedad de los datos. Fue desarrollada por David Dickey y Wayne Fuller. Esta prueba se utiliza para determinar si una serie temporal tiene una raíz unitaria, lo que indica que la serie no es estacionaria. La hipótesis nula de la prueba asume la presencia de una raíz unitaria, y si la prueba rechaza esta hipótesis, se considera que la serie es estacionaria en su forma diferenciada. La estacionariedad es un requisito importante en la modelización econométrica y de series temporales, ya que muchas técnicas asumen que los datos son estacionarios para ser válidas y efectivas." (Tsay, R. S.3rd ed. Wiley., 2010)

2.2.11 Cadena de suministro

La cadena de suministro se puede definir cómo “un conjunto de actividades funcionales, que se repite muchas veces a lo largo del canal de flujo, mediante las cuales la materia prima se convierte en productos terminados y se añade valor para el consumidor”. (Ballou, R., 2004)

Otra definición que se comparte es “la cadena de suministro es el conjunto de actividades involucradas en la gestión del flujo de bienes y servicios, desde la adquisición de materias primas hasta la entrega del producto final al cliente”. (Christopher, M. Pearson UK., 2016)

“La cadena de suministro es el conjunto de procesos que comienzan con la adquisición de materias primas y otros insumos y terminan con la entrega del producto final al cliente, incluyendo la planificación, la gestión de inventarios, la producción y la distribución”. (Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E, McGraw-Hill., 2008)

2.2.12 Pronóstico de la demanda

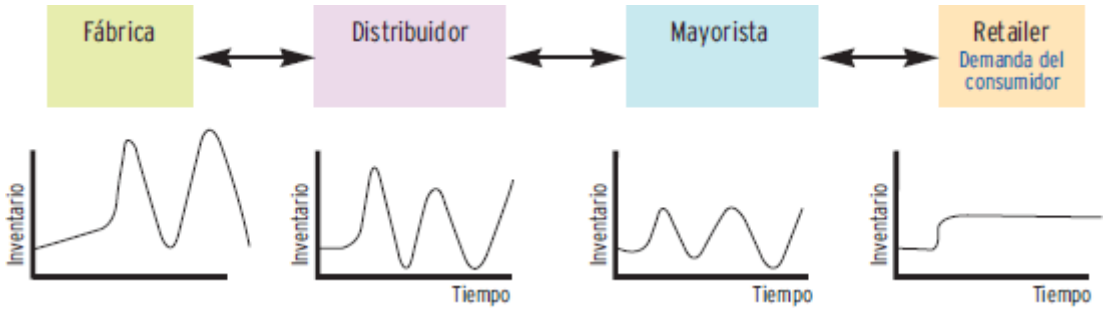
“El pronóstico de la demanda es una parte fundamental de la planificación y la gestión de la cadena de suministro, ya que permite a las organizaciones anticipar y prepararse para las necesidades futuras de los clientes. Se basa en el análisis de datos históricos, tendencias del mercado y otros factores relevantes para predecir con precisión la cantidad de productos o servicios que se requerirán en un período determinado”. (Chopra, S., & Meindl, P. Pearson., 2015)

2.2.13 Efecto látigo

“El efecto látigo es un fenómeno en el que pequeñas fluctuaciones en la demanda de un producto se amplifican a medida que avanza en la cadena de suministro. Mientras que los minoristas realizan pedidos en función de la demanda actual, los mayoristas deben ajustar los pedidos para satisfacer las necesidades de los minoristas, lo que obliga a los fabricantes a ajustar los pedidos para satisfacer las necesidades de los clientes. Este efecto se amplifica a medida que las personas ascienden en la cadena de suministro, lo que puede provocar un exceso o una disminución de la demanda en diferentes niveles de la cadena. El efecto látigo puede resultar costoso para una empresa, ya que puede generar exceso de inventario, costos de inventario y obsolescencia o incluso falta de existencias y pérdida de ingresos”. (Fisher, M.L., 1997, p 75, 105-116)

Figura 17

Efecto látigo



Nota. Adaptado de ¿Qué es efecto látigo? (p. 20), por Fisher, M, 199

2.2.14 Lag Feature

El "lag feature" (característica de rezago) en el contexto del análisis de series de tiempo se refiere a la inclusión de valores pasados de una variable como características adicionales en un modelo predictivo.

Un Lag feature, se refiere a la práctica de utilizar valores pasados de una variable en un conjunto de datos como características predictoras para predecir el valor actual o futuro de esa misma variable. En esencia, un "lag feature" es una representación de una variable en un punto anterior en el tiempo. (Douglas C. Montgomery, Cheryl., 2008)

"Las características de rezago, o la incorporación de valores pasados de una variable como características adicionales, se han utilizado ampliamente en el análisis y pronóstico de series de tiempo. Al incluir valores rezagados como predictores, los modelos pueden capturar dependencias temporales y patrones en los datos, mejorando la precisión de las predicciones". (Hyndman, R. J., & Athanasopoulos, G., 2018)

"El uso de características de rezago en el análisis de series de tiempo es una técnica comúnmente empleada para capturar la autocorrelación y las dependencias temporales en los datos. Al incluir valores pasados como características, se puede mejorar la capacidad predictiva de los modelos en la estimación de valores futuros". (Chatfield, C, London., 2000)

2.2.15 Rolling-Window

Una ventana deslizante, o rolling window, es una técnica utilizada en el análisis de series de tiempo donde se divide la secuencia de datos en subconjuntos de tamaño fijo que se

desplazan a lo largo del tiempo. Esto permite realizar cálculos y análisis en ventanas consecutivas, capturando patrones y tendencias cambiantes en los datos a medida que se desplazan. Es una herramienta útil para obtener una visión dinámica y actualizada de las características de la serie de tiempo.

Rolling-window es una técnica que implica la selección de un intervalo de tiempo fijo y luego se desliza (avanza) a lo largo de la serie temporal para realizar análisis en múltiples ventanas superpuestas.” Esto es útil para examinar y modelar cambios en los datos a lo largo del tiempo y para calcular estadísticas en intervalos específicos” (Hyndman, R.J., & Athanasopoulos, G, 2021)

"Una ventana deslizante, también conocida como ventana móvil, es una técnica utilizada en el análisis de series de tiempo para dividir los datos en subconjuntos de tamaño fijo que se desplazan a lo largo de la secuencia. Esto permite realizar cálculos o análisis en ventanas de tiempo consecutivas y capturar patrones cambiantes en los datos a medida que se mueven en el tiempo" (Shumway, R. H., & Stoffer, D. Springer., 2017)

"La técnica de ventana deslizante es ampliamente utilizada en el análisis de series de tiempo para realizar cálculos y análisis en subconjuntos de datos de tamaño fijo que se desplazan a lo largo de la secuencia temporal. Esto permite capturar tendencias, patrones y cambios en los datos a medida que se mueven en el tiempo, proporcionando una visión dinámica y actualizada de las características de la serie de tiempo". (Brockwell, P.J. & Davis, R. A, Springer., 2016)

2.2.16 Método de correlación Spearman

“Es una herramienta que se utiliza cuando el investigador desea examinar cómo se relacionan o difieren entre sí las variables y los individuos en un estudio. Esta técnica bivariada proporciona representaciones visuales de la información que permiten identificar similitudes o diferencias entre las variables y las personas involucradas.” (Barrera, M, 2014)

“El coeficiente de correlación de Spearman es una medida que evalúa la relación entre dos variables utilizando los rangos u órdenes de los sujetos en cada grupo y comparándolos. Esta medida es especialmente útil cuando se trabaja con un número pequeño de pares de sujetos

(menos de 30). Además de proporcionar información sobre el grado de asociación entre las variables, el coeficiente de Spearman permite determinar si las dos variables aleatorias son dependientes o independientes" (Elorza & Medina Sandoval, 1999, pág. 100).

"El coeficiente de correlación de rangos de Spearman puede puntuar desde -1.0 hasta +1.0, y se interpreta así: los valores cercanos a +1.0, indican que existe una fuerte asociación entre las clasificaciones, o sea que a medida que aumenta un rango el otro también aumenta; los valores cercanos a -1.0 señalan que hay una fuerte asociación negativa entre las clasificaciones, es decir que, al aumentar un rango, el otro decrece. Cuando el valor es 0.0, no hay correlación." (Anderson et al., 1999).

Tabla 8

Grado de relación según coeficiente de correlación

RANGO	RELACIÓN
-0.91 a -1.00	Correlación negativa perfecta
-0.76 a -0.90	Correlación negativa muy fuerte
-0.51 a -0.75	Correlación negativa considerable
-0.11 a -0.50	Correlación negativa media
-0.01 a -0.10	Correlación negativa débil
0.00	No existe correlación
+0.01 a +0.10	Correlación positiva débil
+0.11 a +0.50	Correlación positiva media
+0.51 a +0.75	Correlación positiva considerable
+0.76 a +0.90	Correlación positiva muy fuerte
+0.91 a +1.00	Correlación positiva perfecta

Nota. Hernández, S & Fernández, C (1998).

Según Lugon, A: "Muchos problemas de aprendizaje automático requieren un marco de datos sin variables altamente correlacionadas. La eliminación de estas variables es una técnica simple y efectiva para mejorar el rendimiento de los algoritmos de aprendizaje automático." (2023)

CAPÍTULO III: ENTORNO EMPRESARIAL

3.1 Descripción de la empresa

3.1.1 Reseña histórica y actividad económica

La compañía CBC Peruana S.A.C, con sede en Perú, inició sus actividades en el año 2016 y es parte de una empresa multinacional que distribuye una gran variedad de productos bebibles en la región. A lo largo de los últimos 100 años, ha tenido un papel importante en la industria. De la información obtenida en su página web oficial, la empresa fue fundada en el año 1885 por Enrique Castillo, quien fundó una primera planta embotelladora en Guatemala. Con su ambición y dedicación al trabajo, logró prosperar en las siguientes décadas. En el año 2016, la empresa alcanzó un logro distintivo, se convirtió en una marca de alcance internacional, ajustándose velozmente al mercado de bebidas nutritivas.

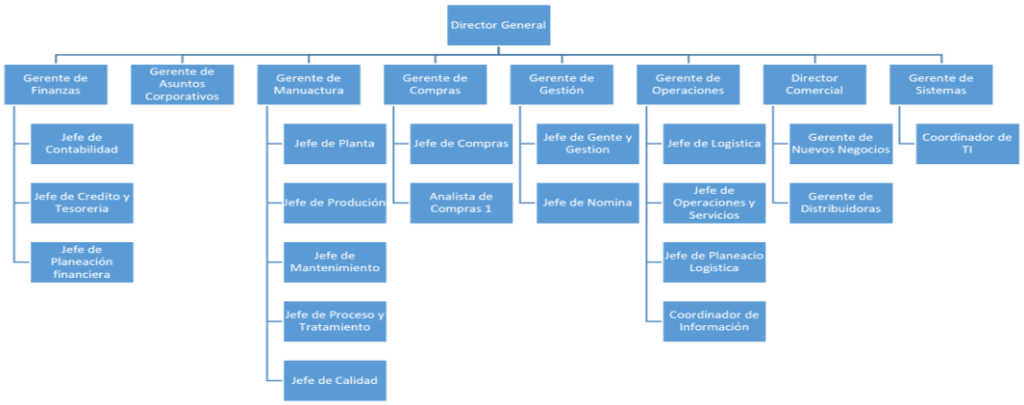
3.2.1 Descripción de la organización

3.2.1.1 Organigrama

La siguiente figura es la estructura organizativa de CBC Peruana S.A.C.

Figura 18

Organigrama



Nota. En base a los datos de CBC Peruana S.A.C

3.2.1.2 Cadena de suministro

A continuación, se detalla la representación gráfica de la cadena de suministro y sus agentes:

Figura 19

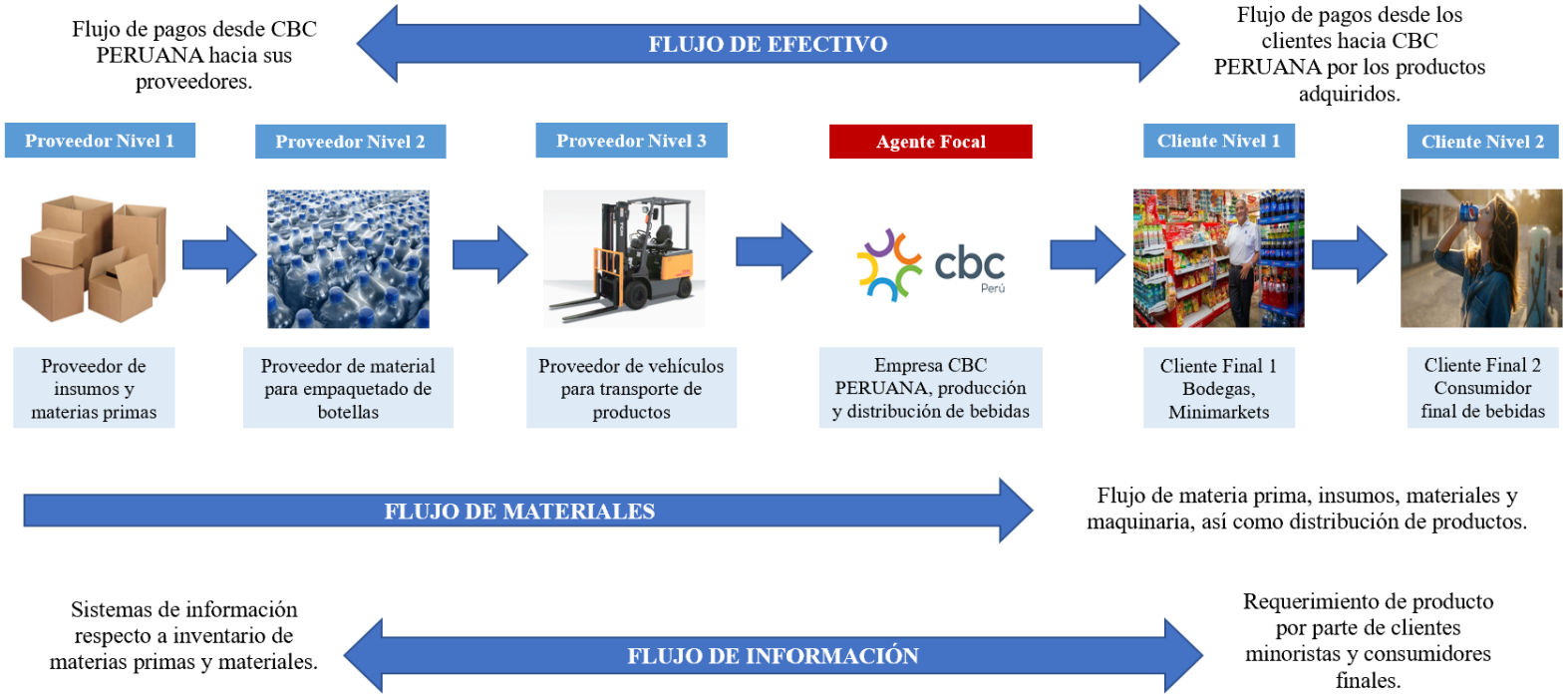
Representación gráfica de la cadena de suministro

Nota. Elaboración propia

Asimismo, se presenta a continuación la ilustración visual de la cadena de suministro y las personas o entidades involucradas en ella.

Figura 20

Representación de la red de suministro



Nota. Elaboración propia

3.3.1 Datos generales estratégicos de la empresa

3.3.1.1 Visión, misión y valores o principios

Visión

"Ser la mejor solución para nuestros clientes, convirtiéndonos en la primera opción de compra en el punto de venta y promoviendo el desarrollo de un mundo mejor" (CBC PERUANA S.A.C, 2023).

Misión

"Generar valor a nuestros clientes y consumidores a través de equipos de alto rendimiento con el mejor portafolio de marcas en todas las ocasiones de consumo" (CBC PERUANA S.A.C, 2023).

Valores o Principios

CBC Peruana S.A.C subraya su enfoque operativo basado en los siguientes seis valores: (a) pasión, trabajando con dedicación y velocidad; (b) disciplina, para obtener resultados sostenibles a través del cumplimiento de procedimientos; (c) aspiramos a grandes logros, persiguiendo cualquier objetivo; (d) integridad, corrigiendo nuestras propias acciones; (e) sentido de pertenencia, creando oportunidades; y (f) contar con personas excelentes para lograr un aprendizaje continuo y alcanzar resultados conjuntos.

3.3.1.2 Objetivos estratégicos

La empresa tiene como meta principal convertirse en la embotelladora y distribuidora líder y más rentable tanto en Perú como a nivel mundial. Reconoce que su capital humano es su activo más valioso y se esfuerza por brindarles la libertad necesaria para crecer a un ritmo acorde a su talento, asegurándose de que su trabajo sea adecuadamente recompensado.

Además, la empresa busca aprovechar el potencial de su equipo para fomentar el liderazgo, centrándose en el desarrollo y formación de líderes aún más capacitados para el futuro.

Un componente clave para alcanzar los objetivos estratégicos es su cultura organizacional, que se distingue por no estar totalmente satisfechos con los resultados logrados. Están persuadidos de que la búsqueda incesante de la excelencia ayuda a conservar una ventaja competitiva perdurable.

3.3.1.3 Evaluación interna y externa

Fortalezas:

F1: Distribuidores exclusivos del portafolio de bebidas de PepsiCo.
F2: Adecuada gestión empresarial.
F3: Amplio y diversificado portafolio de productos.
F4: Facilidades operativas en plantas de producción.
F5: Estrecha relación de coordinación y comunicación con proveedores de insumos y materiales claves.
F6: Empleo de sistemas informáticos en procesos claves.

Oportunidades:

O1: Nuevas medidas del equipo de marketing de PepsiCo para fomentar el debut de nuevas variedades de producto y mejoras.
O2: Disponibilidad de reglas y programas para innovar nuevas tecnologías.
O3: Aumento de la cuota de mercado y de la reputación de marca en los lugares de venta a nivel nacional.
O4: Existencia en todas las zonas del país, tanto de forma directa como indirecta.
O5: Reducción del progreso de la economía, debido a los montos que ofrecen los productos contra la competencia.

Debilidades:

D1: Índices adversos de rentabilidad y rendimiento.
D2: Volatilidad en índices de liquidez.
D3: Algunos productos con muy bajos índices de acogida, formatos poco requeridos por los consumidores.

D4: Atención indirecta, a través de distribuidores terceros en muchas regiones del país.

D5: Desaciertos en la proyección de la demanda, insuficiencia o sobrante de existencias en algunas clases de producto en los años pasados.

Amenazas:

A1: Incertidumbre económica del país.

A2: El mínimo crecimiento tecnológico del país, en cuanto al desarrollo de inteligencias artificiales, para elaborar pronósticos.

A3: Planes comerciales hostiles por parte de los más destacados competidores.

A4: Posicionamiento de marcas líderes en la industria de bebidas.

A5: Manifestación próxima del fenómeno del niño en el verano siguiente.

Tabla 9

Matriz factores internos determinantes de éxito CBC peruana S.A.C

Factores internos determinantes de éxito	Peso	Calificación	Peso Ponderado
Fortalezas			
Distribuidores exclusivos del portafolio de bebidas de PepsiCo	8%	4	0.32
Adecuada gestión empresarial.	7%	4	0.28
Amplio y diversificado portafolio de productos.	12%	3	0.36
Facilidades operativas en plantas de producción	9%	3	0.27
Estrecha relación de coordinación y comunicación con proveedores de insumos y materiales claves.	17%	4	0.68
Empleo de sistemas informáticos en procesos claves	11%	3	0.33
Debilidades			
Índices adversos de rentabilidad y rendimiento.	4%	2	0.08
Volatilidad en índices de liquidez.	4%	1	0.04
Algunos productos con muy bajos índices de acogida, formatos poco requeridos por los consumidores	7%	2	0.14
Atención indirecta, a través de distribuidores terceros en diferentes regiones del país.	6%	2	0.12
Desaciertos en la proyección de la demanda, insuficiencia o sobrante de existencias en algunas clases de producto en los años pasados.	15%	2	0.3
Total	**100%**		**2.92**

Nota. Elaboración propia

Tabla 10

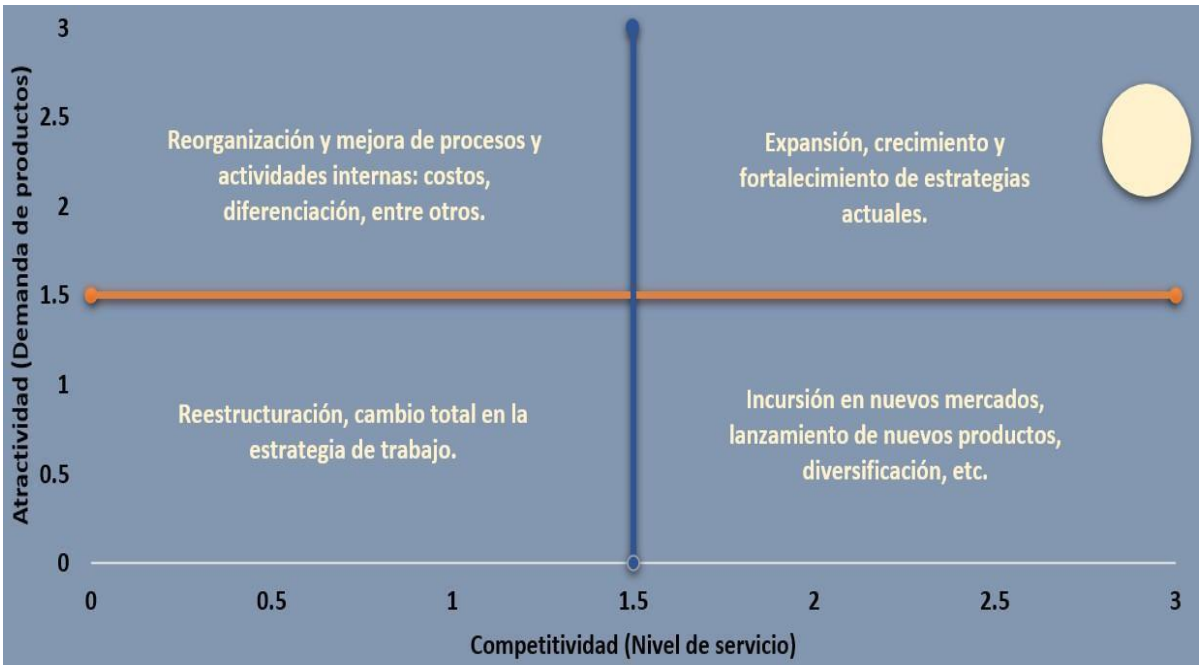
Matriz factores externos determinantes de éxito CBC peruana S.A.C

Factores externos determinantes de éxito	Peso	Calificación	Peso Ponderado
Oportunidades			
Nuevas medidas del equipo de marketing de PepsiCo para fomentar el debut de nuevas variedades de producto y mejoras	9%	4	0.36
Disponibilidad de reglas y programas para innovar nuevas tecnologías	12%	3	0.36
Aumento de la cuota de mercado y de la reputación de marca en los lugares de venta a nivel nacional	8%	4	0.32
Existencia en todas las zonas del país, tanto de forma directa como indirecta	7%	3	0.21
Reducción del progreso de la economía, debido a los montos que ofrecen los productos contra la competencia.	14%	3	0.42
Amenazas			
Incertidumbre económica del país.	14%	1	0.14
El mínimo crecimiento tecnológico del país, en cuanto al desarrollo de inteligencias artificiales, para elaborar pronósticos	9%	1	0.09
Planes comerciales hostiles por parte de los más destacados competidores	12%	2	0.24
Posicionamiento de marcas líderes en la industria de bebidas	7%	1	0.07
Manifestación próxima del fenómeno del niño en el verano siguiente	8%	2	0.16
Total	**100%**		**2.37**

Nota. Elaboración propia.

Figura 21

Análisis de la situación interna y externa



Nota Elaboración propia

El análisis de la situación interna y externa, indica que la empresa CBC Peruana S.A.C se encuentra en el segundo cuadrante. Por lo tanto, la estrategia de implementación debe enfocarse en la expansión, crecimiento y fortalecimiento de las estrategias existentes.

Del análisis de las debilidades de la empresa, se identificó que los errores en los pronósticos de demanda y los productos con baja aceptación son debilidades significativas para la empresa. Por tanto, es necesario poner énfasis en abordar estas debilidades en conjunto con las estrategias de expansión y crecimiento.

En este contexto, la propuesta de implementar técnicas de Machine Learning, brindará a la empresa una ventaja competitiva en el mercado, debido a que contribuirá a reducir los costes de producción e inventario y mejorar la eficiencia financiera de la empresa.

3.4.1 Modelo de negocio actual (CANVAS)

Figura 22

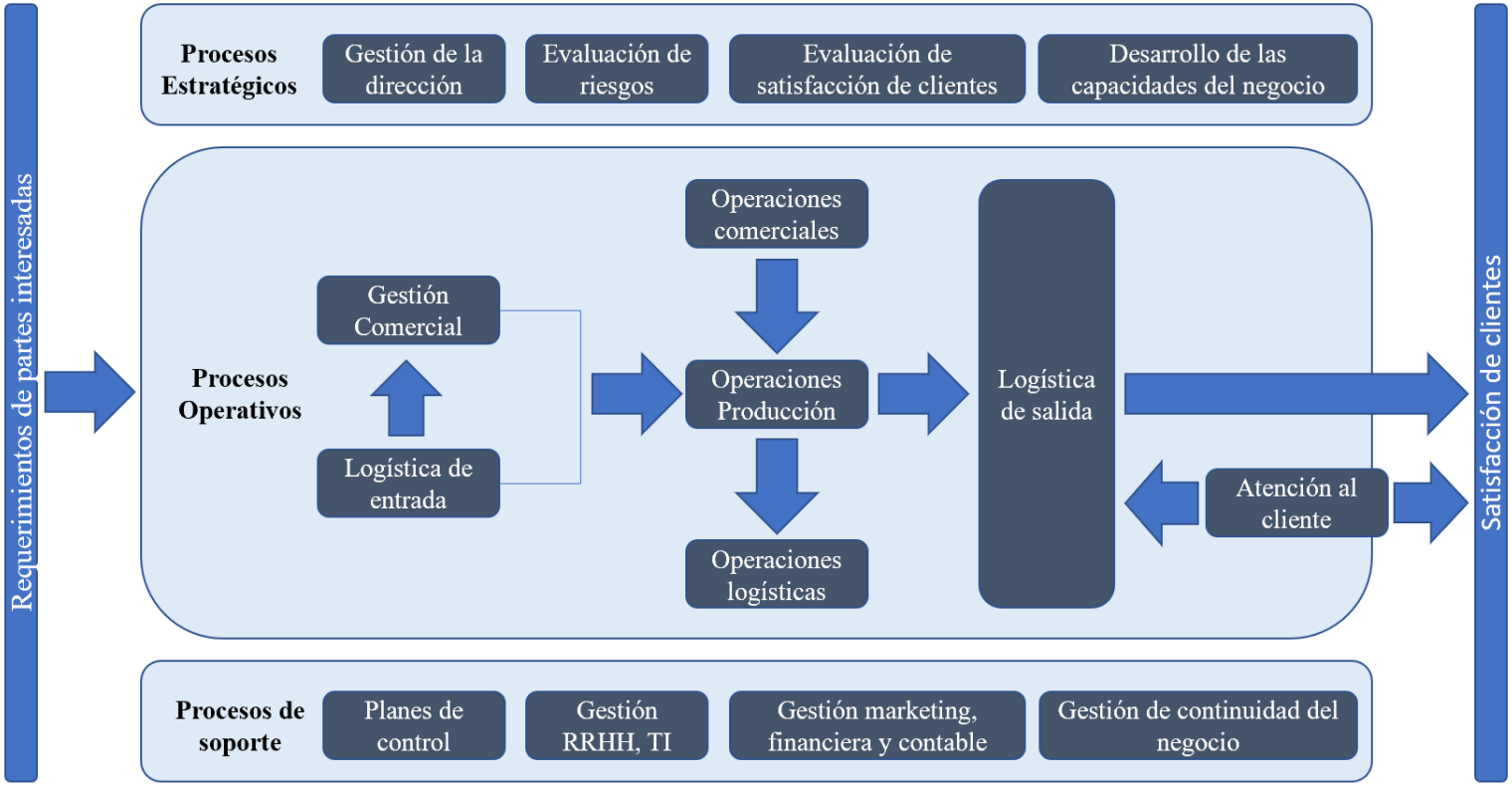
Modelo Canvas

Nota. En base a datos de CBC Peruana S.A.C

3.5.1 Mapa de procesos actual

Figura 23

Mapa de procesos



Nota. Elaboración propia

Adicionalmente, se proporcionará la descripción de los siguientes procesos:

Procesos Estratégicos

- Gestión de la Dirección: Asegurarse de que los líderes y miembros clave de la empresa manejen adecuadamente los procesos del negocio.
- Análisis de riesgos: Evaluar los posibles escenarios que podrían afectar los procesos y los resultados de la empresa.
- Seguimiento de la satisfacción de los clientes: Mantener un monitoreo constante de los indicadores de servicio y el crecimiento de las ventas.
- Desarrollo de las Capacidades del Negocio: Garantizar la implementación efectiva de estrategias para aumentar las ventas y mejorar el enfoque empresarial en todas las áreas de la organización, generando así un mayor compromiso en todo el equipo de trabajo.

Procesos Operativos

- Gestión Comercial: Consiste en liderar las ventas, ofertas y acciones del grupo de ventas, adaptándose a las necesidades del mercado. Asimismo, ofrece apoyo para alcanzar las metas mensuales de ventas y la fidelización de los clientes, todo ello evaluado mediante indicadores de negocio.
- Logística de Entrada: Hace referencia al proceso de abastecimiento o compras realizado por el equipo, basado en cálculos y proyecciones, con el objetivo de asegurar un nivel de inventario seguro y la producción necesaria para satisfacer la demanda del mercado.

Operaciones:

- Operaciones Comerciales: Son las actividades diarias que realiza el equipo de ventas para concretar ventas y ofrecer un servicio de calidad. El enfoque principal es brindar una atención de calidad, aumentar los márgenes y buscar liderazgo en el mercado.
- Operaciones de producción: Esta área se encarga principalmente de fabricar el producto, pero también se ocupa de otras tareas como evaluar los productos

o servicios proporcionados, medir los tiempos de ejecución, asegurar la seguridad e higiene, establecer procedimientos operativos, supervisar la calidad y manejar los niveles de inventario, entre otras funciones.

- Operaciones Logísticas: Engloba todas las actividades relacionadas con las compras, distribución y almacenamiento.
- Logística de Salida: Es la fase de la cadena de suministro que asegura la entrega del producto acabado al cliente último. También abarca la entrega de productos a otras dependencias de la empresa.
- Atención al cliente: El propósito es confirmar la calidad en el suministro de productos a los clientes, siguiendo todos los parámetros definidos. Además, proporciona un conveniente respaldo en caso de solicitudes adicionales o cualquier situación desagradable que pueda ocurrir para los clientes. Es importante tener en cuenta que el tiempo y la precisión en la entrega de los pedidos son indicadores fundamentales para medir la satisfacción del cliente.

Procesos de Soporte:

- Planes de control: Desarrollo de procedimientos operativos estándar, políticas, objetivos, procesos, documentos y recursos con el propósito de gestionar y asegurar la calidad de las operaciones y del producto o servicio ofrecido por la empresa.
- Gestión de RRHH, TI: Asegurarse de adquirir, desarrollar y mantener el capital humano de la empresa, así como asegurar el adecuado desempeño de los sistemas de tecnología e información. También se encarga de explorar opciones de desarrollo tecnológico.
- Gestión de Marketing, financiera y contable: Elaboración de estimaciones sobre la repercusión de la comunicación de la marca con el público propósito, así como análisis de la rentabilidad del negocio conforme a las acciones elaboradas y el presupuesto destinado a cada área de la empresa.
- Gestión de Continuidad del Negocio: Ejecución de acciones y medidas para asegurar la continuidad sin interrupciones de las operaciones clave del negocio.

CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

4.1 Diseño de la Investigación

4.1.1 Enfoque de la investigación

La presente investigación, adopta un enfoque cuantitativo, para desarrollar herramientas que faciliten la identificación y medición de atributos específicos relacionados con la demanda de las bebidas gaseosas Concordia de Piña de 03 litros en paquetes de 04 unidades y Evervess Ginger de 1.5 litros en paquetes de 06 unidades. Para ello, se utilizarán técnicas estadísticas, modelado de datos, mediante los enfoques de Forecasting y Regresión, que permitirán una planificación más eficiente en la predicción de la demanda. Por consiguiente, contribuirá en la elección del número de recursos imprescindibles, como insumos, personal y maquinaria, para cubrir la demanda y prevenir inconvenientes de exceso o falta de existencias.

4.1.2 Alcance de la Investigación

La investigación aplica un análisis correlacional, con la finalidad de examinar la relación entre las variables y realizar un análisis de pronóstico para predecir los valores futuros de la variable objetivo, y su relación a lo largo del tiempo. En este sentido, se requiere predecir la cantidad de paquetes vendidos de las bebidas Concordia de Piña 03 litros de 04 unidades y Evervest Ginger de 1.5 litros de 06 unidades, en intervalos de tiempo de 03, 06 y 12 meses para el año 2024. La recopilación de datos para el pronóstico se basará en la obtención de registros históricos de la demanda, desde enero del 2019 hasta julio del 2023. Para ello, se utilizarán técnicas de Machine Learning, incluyendo Regresión Lineal, LightGBM Regressor y series de tiempo como SARIMA y FB Prophet, bajo los enfoques de Forecasting y Regresión

4.1.3 Tipo de la investigación

La presente investigación aborda un diseño experimental, debido a que, implica manipular una o más variables para observar su efecto en otra variable y posteriormente, efectuar las predicciones de la demanda. Para series de tiempo, la investigación experimental implica realizar pruebas con diferentes modelos, hiper parámetros y técnicas de preprocesamiento de datos para determinar cómo afectan a la precisión de las predicciones.

4.1.4 Población y Muestra

El detalle de la población y muestra se expone a continuación:

Tabla 11

Población y muestra

Población	Muestra
Cantidad de bebidas vendidas desde el mes de enero del año 2019 al mes de julio del año 2023	Cantidad de paquetes vendidos de la bebida Concordia de Piña de 03 litros de 04 unidades, desde el 01 enero del 2019 al 31 de julio del 2023.
	Cantidades de paquetes vendidos de la bebida Evervess Ginger de 1.5 litros de 06 unidades desde el 01 enero del 2019 al 31 de julio del 2023.

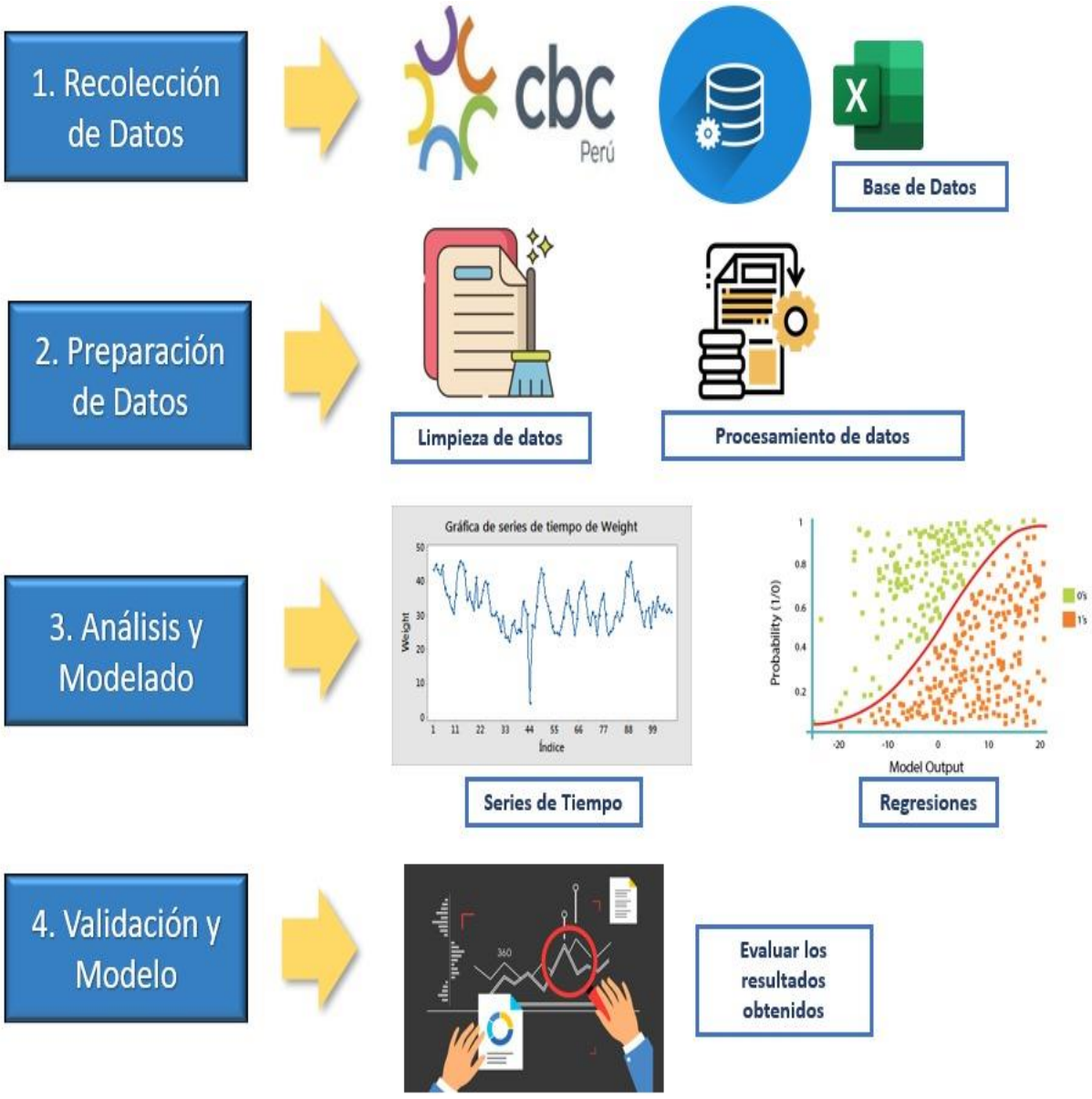
Nota. Elaboración propia

4.2 Metodología de implementación de la solución

Esta investigación, emplea la metodología Industry Standard Process for Data Mining, la cual es ampliamente reconocida y utilizada en el análisis de datos. Dicha metodología consta de cuatro etapas y se presenta en detalle en la Figura 24.

Figura 24

Metodología de implementación



Nota. Elaboración Propia

4.2.1 Recopilación de Datos

Para este estudio, se recopilaron los datos de la empresa CBC Peruana S.A.C. Estos datos incluyen información histórica de la cantidad vendida de los productos Concordia de Piña 03 litros en paquetes de 04 unidades y Evervess Ginger de 1.5 litros en paquetes de 06 unidades, desde enero de 2019 hasta julio de 2023.

4.2.2 Preparación de datos

En esta etapa, se explorará la información obtenida para comprender la composición de los datos, incluyendo la identificación de patrones estacionales. Luego, se llevará a cabo un preprocesamiento de los datos para eliminar registros duplicados, datos ausentes o incorrectos, valores atípicos (outliers) y otros ajustes necesarios.

4.2.3 Análisis y modelado

En esta fase, se construirán cuatro modelos de Machine Learning supervisado: Regresión Lineal, LightGBM Regressor, SARIMA y FB Prophet. Se utilizará lenguaje de programación Python y la plataforma web Colab. Para la construcción de los modelos, se aplicará estandarización y transformación logarítmica a los datos.

4.2.4 Validación de modelo

La validación del modelo consiste en evaluar los resultados de los modelos de Machine Learning. Para ello, se utilizará las métricas RMSE (Root Mean Square Error), MAE (Mean Absolute Error) y MAPE (Mean Absolute Percentage Error), con el objetivo obtener el modelo que mejor se ajuste a los datos. El modelo seleccionado, se utilizará para predecir la cantidad de paquetes vendidos de la bebida Evervess Ginger de 1.5 litros de 06 unidades y la bebida Concordia de Piña de 03 litros de 04 unidades, en periodos de tiempo 03, 06 y 12 meses para el año 2024.

4.3. Metodología para la medición de resultados de la implementación

Tabla 12

FB Prophet y SARIMA

Tipo de Variable	Variable	Responsable	Métricas
Independiente	Cantidad de paquetes vendidos de bebida Concordia de Piña de 03 litros de 04 unidades desde enero del 2019 a julio del 2023	Unidad de Producción y Operaciones	MAE: Mean Absolute Error MAPE: Mean Absolute Pecertange Error RSME: Root Mean Square Error
Dependiente	Dimensión de tiempo desde el mes de enero del 2019 al mes de julio del 2023	Unidad de Producción y Operaciones	
Independiente	Cantidad de paquetes vendidos de bebida Evervess Ginger de 1.5 litros de 06 unidades desde enero del 2019 a julio del 2023	Unidad de Producción y Operaciones	MAE: Mean Absolute Error MAPE: Mean Absolute Pecertange Error RSME: Root Mean Square Error
Dependiente	Dimensión de tiempo desde el mes de enero del 2019 al mes de julio del 2023	Unidad de Producción y Operaciones	

Nota. Elaboración propia

Tabla 13

Regresión Lineal y LGBM Regressor

Tipo de Variable	Variable	Responsable	Métricas
Independiente	Cantidad de paquetes vendidos de bebida Concordia de Piña de 03 litros de 04 unidades desde enero del 2019 a julio del 2023	Unidad de Producción y Operaciones	MAE: Mean Absolute Error MAPE: Mean Absolute Percentage Error RSME: Root Mean Square Error
Dependiente	Dimensión de tiempo desde el mes de enero del 2019 al mes de julio del 2023. *Tipo de atención de venta. *Regiones de Venta.	Unidad de Producción y Operaciones	
Independiente	Cantidad de paquetes vendidos de bebida Evervess Ginger de 1.5 litros de 06 unidades desde enero del 2019 a julio del 2023	Unidad de Producción y Operaciones	MAE: Mean Absolute Error MAPE: Mean Absolute Percentage Error RSME: Root Mean Square Error
Dependiente	Dimensión de tiempo desde el mes de enero del 2019 al mes de julio del 2023 *Tipo de atención de venta. *Regiones de Venta.	Unidad de Producción y Operaciones	

**Variables que se utilizarán bajo el enfoque del problema de Regresión.*

Nota. Elaboración propia

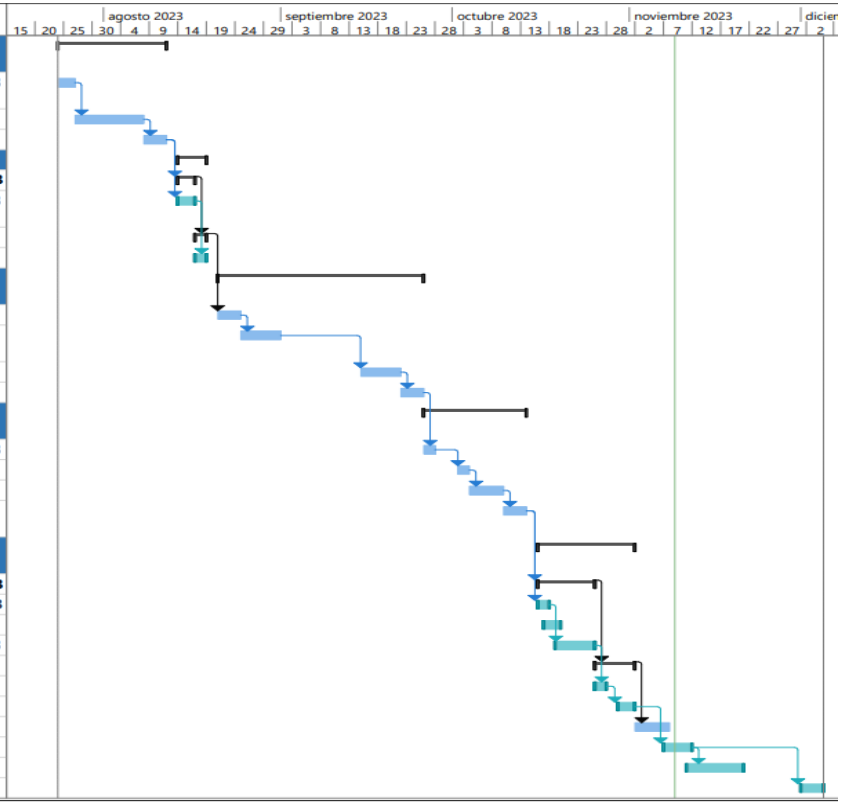
4.4 Cronograma de actividades y presupuesto

4.4.1 Cronograma de Actividades

Figura 25

Cronograma de actividades

Nombre de tarea	Duración	Comienzo	Fin
CAPITULO I : PLANTEAMIENTO DEL PROBLEMA	**15 días**	**lun 24/07/23**	**vie 11/08/23**
Descripción de la Realidad Problemática	3 días	lun 24/07/23	mié 26/07/23
Justificación de la investigación	8 días	jue 27/07/23	lun 7/08/23
Delimitación de la investigación	4 días	mar 8/08/23	vie 11/08/23
CAPITULO II: MARCO TEORICO	**5 días**	**lun 14/08/23**	**vie 18/08/23**
Antecedentes	**3 días**	**lun 14/08/23**	**mié 16/08/23**
Revisión de investigaciónes pasadas	3 días	lun 14/08/23	mié 16/08/23
Bases teóricas	**2 días**	**jue 17/08/23**	**vie 18/08/23**
Recopilación de información	2 días	jue 17/08/23	vie 18/08/23
CAPITULO III :ENTORNO EMPRESARIAL	**26 días**	**lun 21/08/23**	**lun 25/09/23**
Descripción de la empresa	4 días	lun 21/08/23	jue 24/08/23
Datos generales estratégicos de la empresa	5 días	vie 25/08/23	jue 31/08/23
Modelo de negocio actual	5 días	vie 15/09/23	jue 21/09/23
Mapa proceso actual	2 días	vie 22/09/23	lun 25/09/23
CAPITULO IV: METODOLOGÍA DE INVESTIGACIÓN	**14 días**	**mar 26/09/23**	**vie 13/10/23**
Diseño de la investigación	2 días	mar 26/09/23	mié 27/09/23
Metodología de implementación	2 días	lun 2/10/23	mar 3/10/23
Metodología para la medición	4 días	mié 4/10/23	lun 9/10/23
Cronograma de actividades y presupuesto	4 días	mar 10/10/23	vie 13/10/23
CAPITULO V: DESARROLLO DE LA SOLUCIÓN	**13 días**	**lun 16/10/23**	**mié 1/11/23**
Propuesta de solución	**8 días**	**lun 16/10/23**	**mié 25/10/23**
Recolección de datos	2 días	lun 16/10/23	mar 17/10/23
Pre-procesamiento	3 días	mar 17/10/23	jue 19/10/23
Modelamiento	5 días	jue 19/10/23	mié 25/10/23
Medición de la solución	**5 días**	**jue 26/10/23**	**mié 1/11/23**
Validación de modelo	2 días	jue 26/10/23	vie 27/10/23
Resultados de la solución	3 días	lun 30/10/23	mié 1/11/23
Conclusiones y recomendaciones	4 días	jue 2/11/23	mar 7/11/23
Entrega al asesor de TSP	5 días	mar 7/11/23	sáb 11/11/23
Entrega para sustentación de TSP	7 días	sáb 11/11/23	lun 20/11/23
Sustentación del TSP	2 días	vie 1/12/23	lun 4/12/23



Nota. Elaboración propia

4.4.2 Presupuesto

Tabla 14

Presupuesto

Item	Cantidad	Importe por unidades (soles)	Semanas	Total (soles)
Maquinarias y equipos				
Laptops	5	4,500	8	22,500
Mouse	5	60	8	300
Servidor Dell	1	8,000	8	8,000
Anaconda (Python)	1	1,000	8	1,000
Licencia AZURE	1	1,500	8	1,500
Mano de obra				
Desarrollador de Python	3	4,500	8	13,500
Bachiller	2	3,500	8	7,000
Servicios				
Luz	1	500	8	500
Internet	1	300	8	300
		Total		**54,600**

Nota. Elaboración propia

CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN

5.1 Propuesta solución

El presente estudio se enfoca en la resolución de los desafíos existentes en la planificación de la demanda en la empresa CBC Peruana S.A.C, mediante la aplicación de modelos de Machine Learning. El objetivo principal de nuestra investigación consiste en implementar una solución que se fundamentará en la aplicación de modelos de Machine Learning bajo los enfoques Regresión y Forecasting. Esto, con el objetivo de optimizar la planificación de la demanda y mejorar la eficiencia de la cadena de suministro de la empresa.

Para el desarrollo de la solución, la exploración y organización de los datos se llevará a cabo mediante Microsoft Excel. Asimismo, la implementación de los modelos de Machine Learning, se realizará en la plataforma web Colab, empleando lenguaje de programación Python.

5.1.1 Planteamiento y descripción de Actividades

La secuencia de actividades inicia con la recolección de datos, seguida de la preparación de datos, esto, comprende la exploración y preprocesamiento de los datos. Luego, continúa con el modelamiento de la información, y finalmente, la evaluación de las métricas definidas. Este proceso de evaluación proporcionará la base necesaria para seleccionar el modelo más efectivo.

5.1.1.1 Adquisición de datos

La obtención o adquisición de datos, implica la recopilación de información de diferentes fuentes con el objetivo de obtener una visión completa y precisa de los productos específicos. En el contexto de nuestro proyecto de investigación, será esencial recopilar datos históricos de la cantidad de paquetes vendidos de las bebidas Concordia de 03 litros de 04 unidades y Evervess Gin de 1.5 litros de 06 unidades desde el 01 enero del 2019 al 31 de julio del 2023.

5.1.1.2 Preparación

Exploración. Esta fase desempeña un papel crucial en la comprensión de la estructura de los datos, evaluar su comportamiento y realizar un análisis inicial de la información obtenida. En el marco de este proyecto, es imperativo llevar a cabo un análisis del patrón de demanda con el objetivo de identificar el modelo más adecuado para prever el pico máximo de demanda de los productos. Es necesario realizar un análisis en diferentes escalas de tiempo con el fin de comprender el comportamiento de la demanda en cada periodo de tiempo.

Preprocesamiento. La fase de preprocesamiento implica llevar a cabo tareas de depuración de datos, integración de información, ajuste de valores y la creación del conjunto de datos definitivo que se utilizará en la siguiente etapa de modelado. En este contexto de análisis, será necesario formatear las fechas de manera que sean reconocibles por el software. Además, se llevarán a cabo comprobaciones para verificar la integridad de los datos; en caso de que existan campos vacíos, se procederá a completarlos para garantizar la consistencia y confiabilidad del conjunto de datos.

5.1.1.3 Modelamiento

La etapa de modelamiento adquiere relevancia, ya que, en esta etapa se establece el modelo definitivo que permitirá realizar predicciones de la demanda con el menor error de predicción posible. La elección de una metodología inadecuada podría dar lugar a un error sustancial y, en consecuencia, a un modelo que no aporte valor a la investigación. En esta etapa se determinará las técnicas utilizadas, el diseño, construcción y evaluación del modelo. En el marco de esta investigación, se llevará a cabo una evaluación exhaustiva para determinar si los modelos de Regresión Lineal, LightGBM Regressor, y series temporales como SARIMA y FB Prophet son apropiados y se ajustan al comportamiento de la demanda.

5.1.1.4 Evaluación del modelo

En esta fase, se procede a realizar la evaluación de los resultados obtenidos, y se determina la idoneidad del modelo para su implementación en la empresa, y su relevancia en el proceso en el proceso de toma de decisiones de la organización. Se llevará a cabo un análisis

minucioso para identificar si los modelos de Regresión Lineal, LightGBM Regressor o métodos basados en series temporales, como SARIMA y FB Prophet, se ajustan de manera óptima al comportamiento de la demanda y permite obtener una predicción más precisa de los datos.

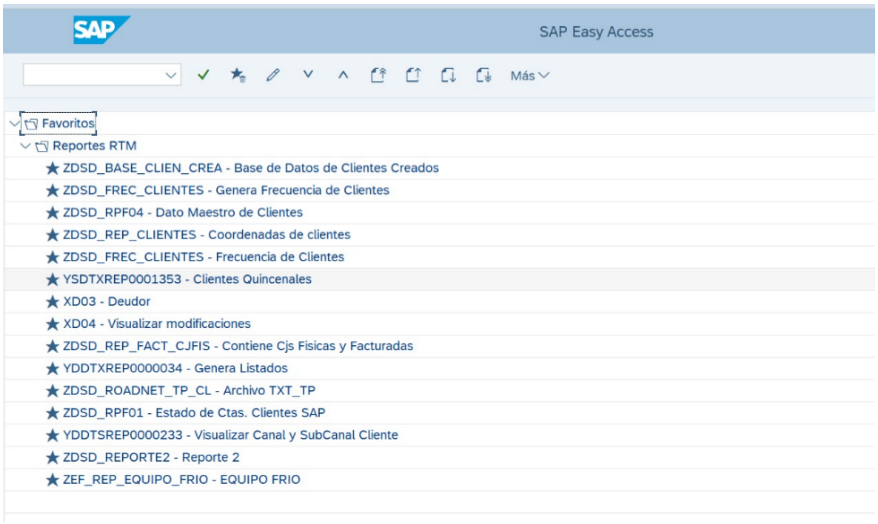
5.1.2 Desarrollo de actividades. Aplicación de herramientas de solución

5.1.2.1 Adquisición de datos

La empresa CBC Peruana S.A.C, gestiona la información de la venta de sus productos, mediante un sistema ERP SAP. Posteriormente, esta información, se exporta en archivos de extensión ".xlsx" de Microsoft Excel. Luego, mediante el servicio de almacenamiento web basado en la nube de Microsoft (SharePoint), se comparte con las diferentes unidades de negocio. A continuación, se presentan los módulos del sistema ERP SAP y repositorios SharePoint de Microsoft.

Figura 26

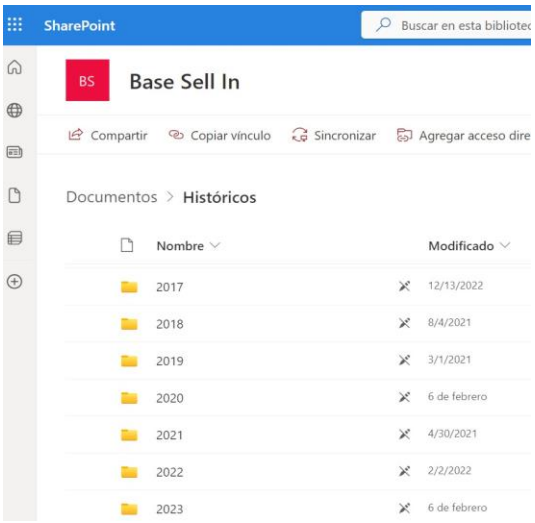
Sistema ERP SAP



Nota. En base de datos de CBC Peruana S.A.C – SAP

Figura 27

Repositorio de OneDrive



Nota: En base a los datos de Repositorio OneDrive

En este sentido, se obtuvo información histórica de las bebidas Concordia de Piña 03 litros de 04 unidades y Evervess Ginger 1.5 litros de 06 unidades, desde el mes de enero del año 2019 hasta al mes de julio del año 2023, en archivos de extensión “.xlsx” y “.xlsb” de Microsoft Excel.

Figura 28

Data mensual

Nombre	Tamaño	Tipo	Fecha de modificación
A – E			
Base Sell In País I Cierre ABR.xlsx	165,304 KB	Hoja de cálculo de Microsoft Excel	8/05/2023 10:23
Base Sell In País I Cierre ENE.xlsb	115,712 KB	Hoja de cálculo binaria de Microsoft Excel	13/10/2023 17:04
Base Sell In País I Cierre FEB.xlsx	176,331 KB	Hoja de cálculo de Microsoft Excel	13/10/2023 17:04
Base Sell In País I Cierre JUL.xlsb	32,053 KB	Hoja de cálculo binaria de Microsoft Excel	13/10/2023 17:04
Base Sell In País I Cierre JUN.xlsx	171,345 KB	Hoja de cálculo de Microsoft Excel	13/10/2023 17:05
Base Sell In País I Cierre MAR I.xlsx	135,370 KB	Hoja de cálculo de Microsoft Excel	13/10/2023 17:05
Base Sell In País I Cierre MAR II.xlsx	79,206 KB	Hoja de cálculo de Microsoft Excel	13/10/2023 17:06
Base Sell In País I Cierre MAY.xlsx	165,358 KB	Hoja de cálculo de Microsoft Excel	13/10/2023 17:06

Nota. En base a la información de CBC Peruana

De la verificación de los datos obtenidos, se identificó, que la bebida Evervess Ginger 1.5 litros de 06 unidades, registra código de producto “BA003740” y asciende a 753,220 registros, mientras que la bebida Concordia de Piña 03 litros de 04 unidades, registra código de producto “BA003727”, y tiene 503,743 registros.

Tabla 15

Datos de bebida Evervess Ginger 1.5 litros de 06 unidades

G. VENTA	ZV	CÓDIGO CLIENTE	CLIENTE	FECHA FACTURA	CÓDIGO SKU	SKU	TOTAL CP	NUEVO NETO	UNIDADES	PAQ C/C	CF C/C	BOZ	DÍA	MES	AÑO	NEGOCIO	SOCIO	UNID. DIAGEO	PAQ TRACKING BI
R26	606508	6034691	CORPORACION CARHUACHUCO S.A.C.	4/01/2023	BA003740	EVERV GINGER	46248.25	38998.40	11100	1850	924.9985	2932.31	4 Ene	2023	CSD	Pepsico		11100	1850
R26	606508	6034691	CORPORACION CARHUACHUCO S.A.C.	3/01/2023	BA003740	EVERV GINGER	21999.16	18550.60	5280	880	439.9993	1394.83	3 Ene	2023	CSD	Pepsico		5280	880
R31	638605	6092334	MASS CENTRALIZADO	5/01/2023	BA003740	EVERV GINGER	17351.1	14704.32	3510	585	292.4995	927.24	5 Ene	2023	CSD	Pepsico		3510	585
R30	638602	6148309	CD10 CENTRAL CENCOSUD	4/01/2023	BA003740	EVERV GINGER	15851	13433.06	3300	550	274.9996	871.77	4 Ene	2023	CSD	Pepsico		3300	550
R31	638605	6092334	MASS CENTRALIZADO	4/01/2023	BA003740	EVERV GINGER	12753.8	10808.32	2580	430	214.9997	681.56	4 Ene	2023	CSD	Pepsico		2580	430
R30	638603	6178602	MAKRO AREQUIPA	3/01/2023	BA003740	EVERV GINGER	9138.8	7744.76	2046	341	170.4997	540.50	3 Ene	2023	CSD	Pepsico		2046	341
R31	638604	6071864	TOTTUS CENTRALIZADO	4/01/2023	BA003740	EVERV GINGER	7767.36	6582.50	1674	279	139.4998	442.22	4 Ene	2023	CSD	Pepsico		1674	279
R31	638605	6071874	PLAZA VEA CENTRALIZADO PUNTA NEG	4/01/2023	BA003740	EVERV GINGER	8215.82	6962.57	1662	277	138.4998	439.05	4 Ene	2023	CSD	Pepsico		1662	277
R87	646400	6178606	MAKRO TRUJILLO 1	4/01/2023	BA003740	EVERV GINGER	6574.06	5571.24	1542	257	128.4998	407.35	4 Ene	2023	CSD	Pepsico		1542	257
R31	638605	6071870	ECONOMAX CUSCO	4/01/2023	BA003740	EVERV GINGER	6347.24	5379.00	1284	214	106.9998	339.20	4 Ene	2023	CSD	Pepsico		1284	214
R30	638603	6178605	MAKRO ICA	3/01/2023	BA003740	EVERV GINGER	5413.6	4587.80	1212	202	100.9998	320.18	3 Ene	2023	CSD	Pepsico		1212	202
R87	645151	6220352	TABERNA DISTRIBUCIONES SAC	3/01/2023	BA003740	EVERV GINGER	5511	4670.34	1200	200	99.9998	317.01	3 Ene	2023	CSD	Pepsico		1200	200
R25	606513	6093479	WILMER JAVIER TORRES CHAVEZ	4/01/2023	BA003740	EVERV GINGER	4999.8	4154.02	1200	200	99.9998	317.01	4 Ene	2023	CSD	Pepsico		1200	200
R87	645501	6124501	INVERSIONES ALVA-VILLA S.A.C	4/01/2023	BA003740	EVERV GINGER	5511	4578.76	1200	200	99.9998	317.01	4 Ene	2023	CSD	Pepsico		1200	200
R87	646400	6178607	MAKRO TRUJILLO 2	4/01/2023	BA003740	EVERV GINGER	4553.24	3858.67	1068	178	88.9999	282.14	4 Ene	2023	CSD	Pepsico		1068	178
R30	603607	6194529	MAKRO CHICLAYO	4/01/2023	BA003740	EVERV GINGER	2599.2	2202.71	684	114	56.9999	180.69	4 Ene	2023	CSD	Pepsico		684	114
R31	603607	6178597	MAKRO PIURA	4/01/2023	BA003740	EVERV GINGER	2211.6	1874.24	582	97	48.4999	153.75	4 Ene	2023	CSD	Pepsico		582	97
R30	638603	6178599	MAKRO SANTA ANITA	5/01/2023	BA003740	EVERV GINGER	2546	2157.63	570	95	47.4999	150.58	5 Ene	2023	CSD	Pepsico		570	95
R31	638605	6083146	VIVANDA ASIA	3/01/2023	BA003740	EVERV GINGER	2432.12	2061.11	492	82	40.9999	129.97	3 Ene	2023	CSD	Pepsico		492	82
R30	638603	6178598	MAKRO SURCO	5/01/2023	BA003740	EVERV GINGER	2063.6	1748.81	462	77	38.4999	122.05	5 Ene	2023	CSD	Pepsico		462	77
R30	638603	6178604	MAKRO CHINCHA	3/01/2023	BA003740	EVERV GINGER	1902.8	1612.55	426	71	35.4999	112.54	3 Ene	2023	CSD	Pepsico		426	71
R30	638603	6195335	MAKRO HUAYLAS	3/01/2023	BA003740	EVERV GINGER	1768.8	1498.98	396	66	32.9999	104.61	3 Ene	2023	CSD	Pepsico		396	66
R30	638603	6224380	MAKRO CHIMBOTE	3/01/2023	BA003740	EVERV GINGER	1608	1362.71	360	60	30.0000	95.10	3 Ene	2023	CSD	Pepsico		360	60

Nota. En base a la información de CBC Peruana

Tabla 16

Concordia de Piña de 03 litros de 04 unidades

G. VENTA	ZV	CÓDIGO CLIENTE	CLIENTE	FECHA FACTURA	CÓDIGO SKU	SKU	TOTAL CP	NUEVO NETO	UNIDADES	PAQ C/C	CF C/C	BOZ	DÍA	MES	AÑO	NEGOCIO	SOCIO	UNID. DIAGEO	PAQ TRACKING BI	CP
R31	638605	6112268	CD MASS SUR	4/01/2023	BA003727	CONCORDIA PI	18079.2	15321.36	4960	1240	826.6668	2620.59	4 Ene	2023	CSD	Pepsico		4960	1240	
R31	638605	6112269	CD MASS NORTE	3/01/2023	BA003727	CONCORDIA PI	10147.68	8599.73	2784	696	464.0001	1470.91	3 Ene	2023	CSD	Pepsico		2784	696	
R31	638605	6115380	MASS ESTE	3/01/2023	BA003727	CONCORDIA PI	9681.12	8204.34	2656	664	442.6668	1403.28	3 Ene	2023	CSD	Pepsico		2656	664	
R31	638605	6115380	MASS ESTE	4/01/2023	BA003727	CONCORDIA PI	6765.12	5733.16	1856	464	309.3334	980.61	4 Ene	2023	CSD	Pepsico		1856	464	
R31	638605	6112269	CD MASS NORTE	5/01/2023	BA003727	CONCORDIA PI	7292.16	6179.80	1536	384	256.0001	811.54	5 Ene	2023	CSD	Pepsico		1536	384	
R30	638602	6148309	CD10 CENTRAL CENCOSUD	4/01/2023	BA003727	CONCORDIA PI	2657.6	2252.21	604	151	100.6667	319.12	4 Ene	2023	CSD	Pepsico		604	151	
R31	638605	6092334	MASS CENTRALIZADO	5/01/2023	BA003727	CONCORDIA PI	1866.24	1581.56	512	128	85.3334	270.51	5 Ene	2023	CSD	Pepsico		512	128	
R56	639701	6082337	CIA DE DISTRIBUCIONES RICHITOR	3/01/2023	BA003727	CONCORDIA PI	1834.8	1265.27	480	120	80.0000	253.61	3 Ene	2023	CSD	Pepsico		480	120	
R31	638605	6092334	MASS CENTRALIZADO	4/01/2023	BA003727	CONCORDIA PI	1749.6	1482.71	480	120	80.0000	253.61	4 Ene	2023	CSD	Pepsico		480	120	
R31	638604	6071864	TOTTUS CENTRALIZADO	4/01/2023	BA003727	CONCORDIA PI	1876.93	1590.63	452	113	75.3333	238.81	4 Ene	2023	CSD	Pepsico		452	113	
R31	603607	6063243	ECONOMAX SULLANA	4/01/2023	BA003727	CONCORDIA PI	1283.04	1087.32	352	88	58.6667	185.98	4 Ene	2023	CSD	Pepsico		352	88	
R31	638605	6071874	PLAZA VEA CENTRALIZADO PUNTA NEG	5/01/2023	BA003727	CONCORDIA PI	933.12	790.78	256	64	42.6667	135.26	5 Ene	2023	CSD	Pepsico		256	64	
R31	638605	6071870	ECONOMAX CUSCO	4/01/2023	BA003727	CONCORDIA PI	816.48	691.94	224	56	37.3333	118.35	4 Ene	2023	CSD	Pepsico		224	56	
R31	603607	6178597	MAKRO PIURA	4/01/2023	BA003727	CONCORDIA PI	749.7	635.34	196	49	32.6667	103.56	4 Ene	2023	CSD	Pepsico		196	49	
R30	638603	6178602	MAKRO AREQUIPA	3/01/2023	BA003727	CONCORDIA PI	924.5	783.48	172	43	28.6667	90.88	3 Ene	2023	CSD	Pepsico		172	43	
R30	638603	6178168	MAKRO VILLA EL SALVADOR	5/01/2023	BA003727	CONCORDIA PI	860	728.81	160	40	26.6667	84.54	5 Ene	2023	CSD	Pepsico		160	40	
R26	606507	6061883	INVERSIONES Y DISTRIBUCIONES	5/01/2023	BA003727	CONCORDIA PI	804.01	668.01	160	40	26.6667	84.54	5 Ene	2023	CSD	Pepsico		160	40	
R02	603501	6065691	MARTINEZ SERNAQUE HENRY	4/01/2023	BA003727	CONCORDIA PI	579.02	481.08	120	30	20.0000	63.40	4 Ene	2023	CSD	Pepsico		120	30	
R02	603607	6207333	MAXI AHORRO AVIACION	5/01/2023	BA003727	CONCORDIA PI	459	388.98	120	30	20.0000	63.40	5 Ene	2023	CSD	Pepsico		120	30	
R30	638603	6178598	MAKRO SURCO	5/01/2023	BA003727	CONCORDIA PI	645	546.61	120	30	20.0000	63.40	5 Ene	2023	CSD	Pepsico		120	30	
R11	606517	6062366	GOLOSINAS JORGITO SAC	5/01/2023	BA003727	CONCORDIA PI	617.97	513.43	120	30	20.0000	63.40	5 Ene	2023	CSD	Pepsico		120	30	
R30	603607	6194529	MAKRO CHICLAYO	5/01/2023	BA003727	CONCORDIA PI	428.4	363.05	112	28	18.6667	59.17	5 Ene	2023	CSD	Pepsico		112	28	

Nota. En base a la información de CBC Peruana

5.1.2.2. Presentación de Variables

De la revisión de la información, se identificó 26 variables, las cuales se proceden a describir a continuación:

Tabla 17

Variables

Item	Variables	Descripción
1	G. Venta	Código del Supervisor de ventas
2	ZV	Código del Vendedor por zona de ventas
3	Código Cliente	Código de cliente
4	Cliente	Datos del cliente
5	Fecha Factura	Fecha de venta del producto
6	Código SKU	Código de Producto
7	SKU	Descripción o denominación del producto
8	Total	Importe total vendido (soles)
9	Cpag	Condición de pago
10	Nuevo Neto	Importe total de ventas sin IGV
11	PAQ C/C	Cantidad de Paquetes Vendidos
12	Cf C/C	Cajas físicas con cargo
13	8oz	Unidad de medida
14	Día	Fecha de venta del producto (Días)
15	Mes	Fecha de venta del producto (mes)
16	Año	Fecha de venta del producto (año)
17	Negocio	Categoría del Producto
18	Socio	Socio estratégico de la marca del producto
19	Unid. Diageo	Unidades de bebidas alcohólicas
20	Paq Tracking Bi	Identificador de Entrega del Producto
21	Cf Diageo	Cajas físicas para bebidas alcohólica
22	Fecha Fac PepsiCo	Fecha de Facturación establecida por PepsiCo
23	Cód Sku PepsiCo	Código del Producto de PepsiCo
24	Atención	Tipo de atención de entrega del producto
25	9l C/C	Unidad de medida
26	Region	Región de Venta

Nota. En base a la información de CBC Peruana

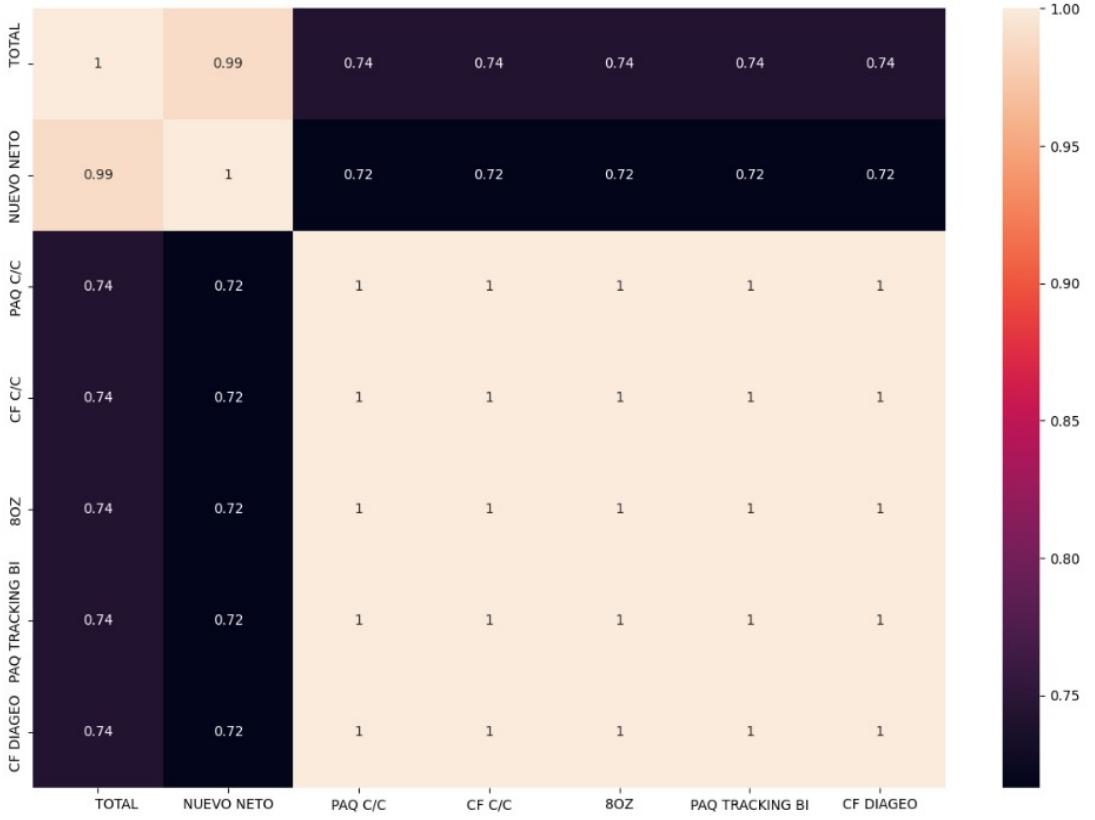
5.1.2.3. Preprocesamiento

En esta etapa, se procedió a evaluar los datos contenidos en cada una de las variables, y se identificó que, de las 26 variables, 16 de ellas con las siguientes características: valores nulos, columnas repetitivas e información irrelevante, las cuales se presentan a continuación: “Zv”, “Código Cliente”, “Cliente”, “Código SKU”, “SKU”, “Cpag”, “Día”, “Mes”, “Año”, “Negocio”, “Socio”, “Unid. Diageo”, “Fecha Fac PepsiCo”, “Cód SKU PepsiCo” y “9l C/C”. Por ende, estas variables no aportan valor a los modelos.

Luego, para las variables numéricas (independientes): “CF Diageo”, “Paq Tracking Bi”, “8oz”, “Cf C/C”, “Nuevo Neto”, “Total”, en contraste con la variable dependiente (“PAQ C/C”), se procedió a realizar una matriz de correlación mediante el método de Spearman, donde se identificó que, las variables se encuentran muy correlacionadas entre sí. Por lo tanto, no serían útiles como variables predictoras en el modelo.

Figura 29

Matriz de Correlación con variables numéricas y la variable Target.



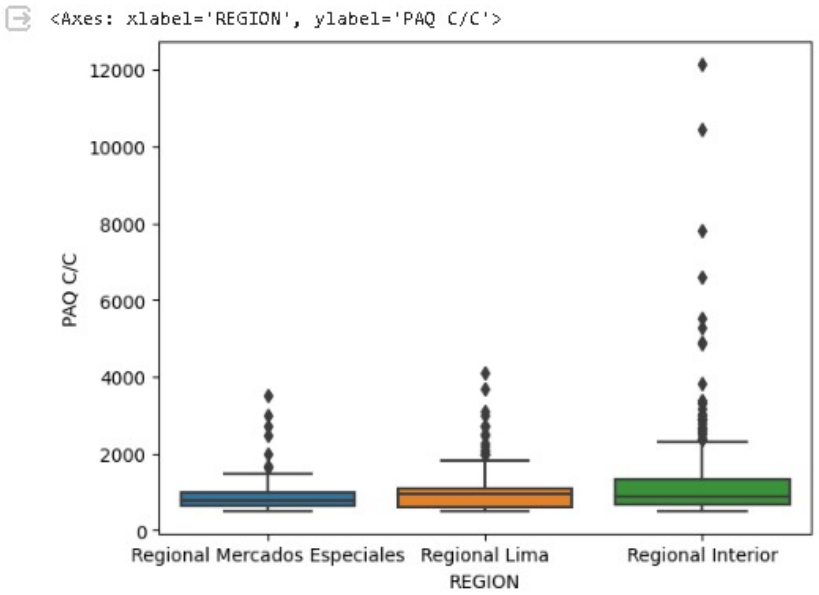
Nota. Elaboración propia

Asimismo, para las variables categóricas: “Region”, “G.Venta”, “Atencion”, se procedió a realizar un análisis exploratorio por medio de gráficos de cajas y bigotes.

Del análisis bivariado entre las variables “Region” y “PAQ C/C”, se identificó que, la mediana de cada una de las clases de la variable “Region”, difiere. Esto, nos indica que, esta variable puede considerarse como una variable predictora que puede aportar valor a los modelos.

Figura 30

Gráfico de cajas y bigotes de las variables 'REGION' y la variable 'PAQC/C'

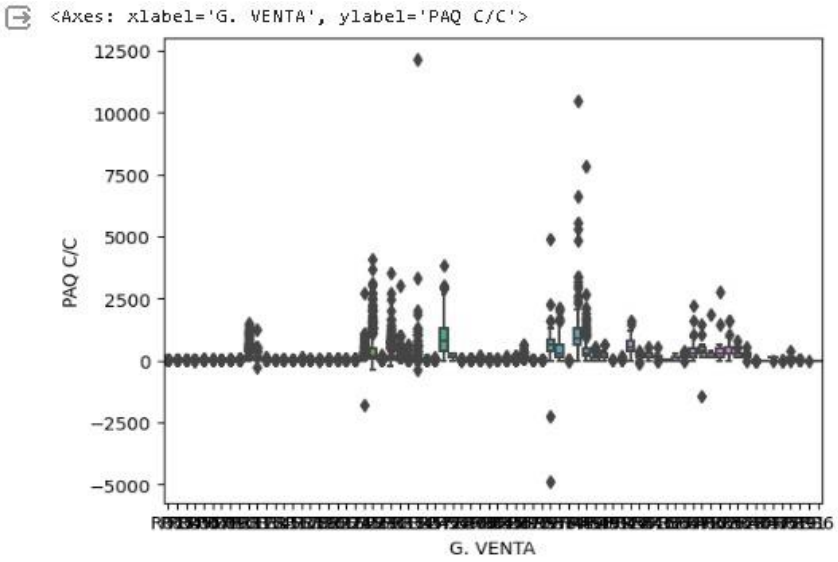


Nota. Elaboración propia

Asimismo, del análisis entre las variables, "G. Venta" y "PAQ C/C", se observa que, las clases de la variable "G. Venta", son muy dispersas, y presentan mayores valores atípicos respecto a la variable "Region". Por lo tanto, esta variable no aporta valor a los modelos.

Figura 31

Gráfico de cajas y bigotes de las variables 'G. Venta' y la variable 'PAQC/C'

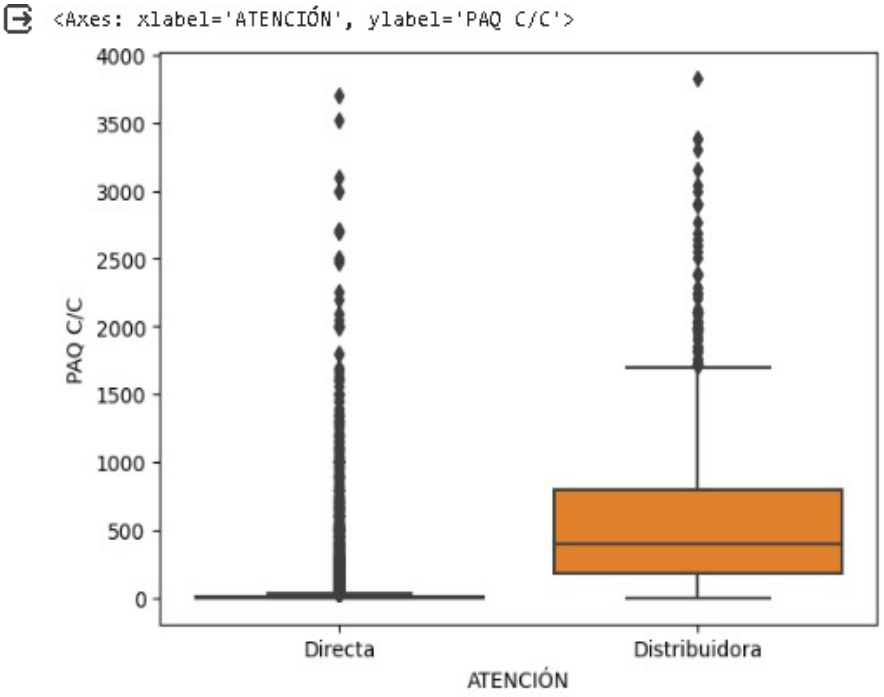


Nota. Elaboración propia

De la relación entre las variables “Atencion” y “PAQ C/C”, se identificó que, la variable “Atención”, tiene un cambio significativo respecto a la variable “PAQ C/C”, lo cual se observa en la diferencia de la mediana para ambas categorías de la variable “Atención”. Por ende, es una variable que aporta valor a los modelos.

Figura 32

Gráfico de cajas y bigotes de las variables ‘Atencion’ y ‘PAQC/C’



Nota. Elaboración propia

Del análisis de la relación entre las variables, se determinó que, para el enfoque de Regresión, se emplearán las variables “Atencion”, “Region” y “Fecha”, mediante la aplicación de los modelos de Machine Learning de Regresión Lineal y LGBM Regressor. Por otro lado, para el enfoque de Forecasting, se utilizarán como valores de entrada y/o variables “Fecha” y “PAQ C/C”, mediante la aplicación de los modelos de Machine Learning de Regresión Lineal, LGBM Regressor y series temporales (SARIMA y FB Prophet). La selección de estas variables se basó en su relevancia y su impacto en la predicción de la variable dependiente.

Enfoque Forecasting

Para el desarrollo del enfoque de Forecasting, se procedió a realizar un análisis del comportamiento histórico de las bebidas Concordia de Piña de 03 litros de 04 unidades y Evervess Ginger de 1.5 litros de 06 unidades durante el periodo de 01 enero del 2019 al 31 de julio del 2023.

Del análisis, se observó que, la cantidad de paquetes vendidos de la bebida Evervess Ginger de 1.5 litros de 06 unidades registra un comportamiento ascendente entre los años 2019, 2020 y 2021, mientras que en el año 2022 se registra un descenso. Sin embargo, en el año 2023 se observa una tendencia al alza. Asimismo, se identificó que registra comportamiento ascendente en los meses de diciembre, periodo que representa fechas festivas, y que puede influir en el comportamiento de la demanda. Por otro lado, en los meses de abril y mayo de cada año, se observa que la cantidad de productos vendidos desciende.

Figura 33

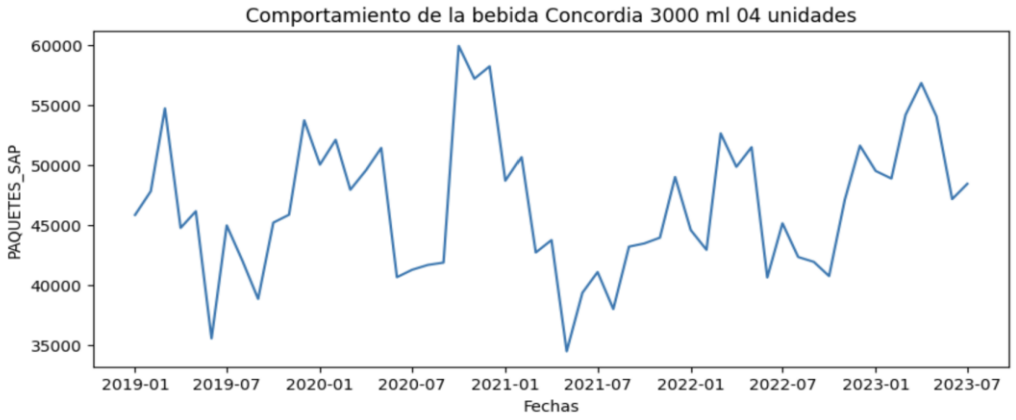
Comportamiento de la venta de la bebida Evervess Ginger

Nota: Elaboración propia

Respecto a la cantidad de paquetes vendidos de bebida Concordia de Piña de 03 litros de 04 unidades, se observa un comportamiento repetitivo de incremento y disminución de ventas entre el mes de enero del 2019 a julio del 2023. Asimismo, se evidencia que, en los meses de febrero, marzo, y julio, se presentan caídas consecutivas.

Figura 34

Comportamiento de venta del producto Concordia de Piña



Nota: Elaboración propia

Luego de analizar el comportamiento de los productos. Se procedió a dividir el conjunto de datos de la bebida Evervess Ginger de 1.5 litros de 06 unidades en el dataframe “datos1”, y en el dataframe “datos2”, el conjunto de datos de la bebida Concordia de Piña de 03 litros de 04 unidades. Posteriormente, se utilizó la función “info” de la librería *Pandas,* con la finalidad de conocer la estructura del dataframe “datos1” y “datos2”, como: tipo de dato, cantidad de registros y verificación de valores nulos. Al respecto, se identificó que, el dataframe “datos1”, contiene 1,483 registros, mientras que el dataframe “datos2”, contiene 1,500 registros. Asimismo, se verificó que, la variable “Fecha Factura”, en ambos dataframe se encuentra en formato de fecha (“datetime”) y no registran valores nulos. Por otro lado, se verificó que, el campo de la cantidad de paquetes (variable dependiente) se encuentra denominado con el nombre “Suma de PAQUETES SAP”, por ende, se renombró a “PAQUETES_SAP” y el campo “FECHA_FACTURA” se renombró a “Fechas”.

Luego, se realizó la preparación de los datos de los dataframe “datos1” y “datos2”. Después, se aplicó formato de año, mes y día (YYYY-MM-DD) al campo “Fechas”, esto, con la finalidad que el software reconozca el tipo de formato. Asimismo, se indexó las fechas, para establecer la frecuencia en días, y completar los días faltantes con datos nulos; luego, se ordenó la información de manera ascendente.

Figura 35

Preparación de datos del dataframe " datos1"

```
[ ]  # Preparación del dato
     # ======================================================================
     datos1['Fechas'] = pd.to_datetime(datos1['Fechas'], format='%Y-%m-%d')
     datos1 = datos1.set_index('Fechas')
     datos1 = datos1.rename(columns={'x': 'y'})
     datos1 = datos1.asfreq('d')
     datos1 = datos1.sort_index()
     datos1.head(4)
```

	PAQUETES_SAP
Fechas	
2019-01-02	2524.0
2019-01-03	2678.0

Nota: Elaboración propia

Luego, para verificar si los datos del campo “Fechas” se encuentran completos en los dataframe “datos1” y “datos2”, se aplicó la función de índice temporal (date_range), cuya función toma el dato más antiguo al más reciente, y brinda como respuesta una expresión booleana “True o False”. En este sentido, “True”, indica si la información está completa y “False” si es que existen datos faltantes. Los resultados de la verificación de índice temporal del campo “Fechas” muestran un valor “True”, lo que significa que los datos en cada dataframe (“datos1” y “datos2”), se encuentran completos.

Asimismo, para el campo "PAQUETES_SAP", se realizó la verificación de valores nulos. Para ello, se aplicó la técnica de Interpolación, mediante la función “interpolate”. Esta técnica se utiliza para rellenar los valores faltantes en un dataframe o serie utilizando diferentes técnicas de interpolación. La función reemplaza los valores nulos o NaN (Not a Number), con valores interpolados que se encuentran entre los valores conocidos. Este procedimiento, se aplicó para los dataframe “datos1” y “datos2”.

Figura 36

Interpolación de datos

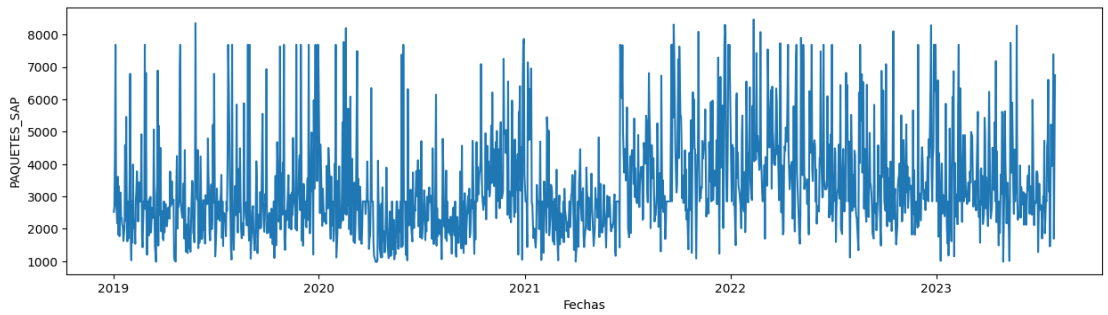
```
# Aplicar la interpolación para completar los valores faltantes
datos1I = datos1.interpolate()
datos2I = datos2.interpolate()
```

Nota: Elaboración propia

Luego, se restableció los índices de los dataframe “datos1” y “datos2”, mediante la función “datetime”, y se obtiene los siguientes gráficos.

Figura 37

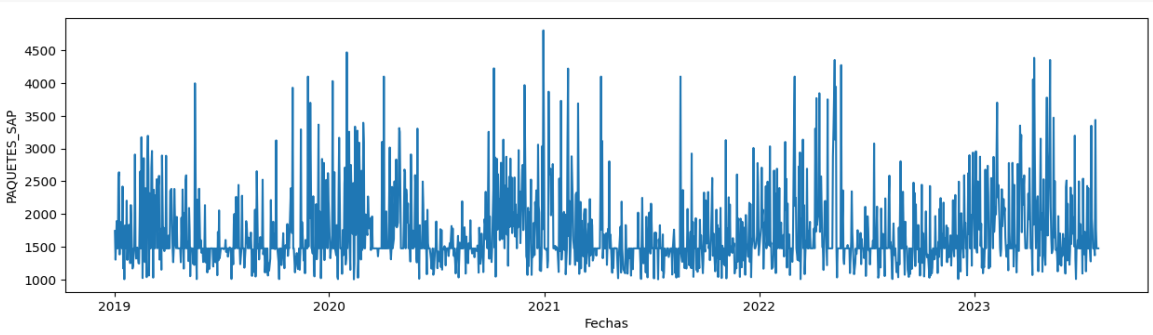
Indexación de Fechas del dataframe “datos1”



Nota: Elaboración propia

Figura 38

Indexación de Fechas del dataframe “datos2”



Nota: Elaboración propia

Las actividades de preprocesamiento descritas en párrafos anteriores son aplicables para los modelos de series temporales: SARIMA y FB Prophet. Sin embargo, para la construcción de los modelos de Regresión Lineal y LightGBM Regressor, es necesario, crear features. Estos features se crearon con Lags y Rolling Windows. Para los dataframe “datos1” y “datos2”, se crearon 30 Lags y 30 Rolling Window, que equivalen a 01 mes (30 días) de desplazamiento.

Para el desarrollo de este procedimiento, se aplicó la función "rolling" y “shift”. La función “rolling”, se utilizó para calcular el método de rolling window en el conjunto de datos, mientras que la función "shift" se utilizó para calcular el método de Lag. Ambas funciones pueden producir valores NaN (Not a Number). Los valores NaN se refiere a los valores faltantes o nulos en un conjunto de datos que se utilizan para calcular la función de lag o rolling window. A partir de esto, se creó los dataframe “lag_features1” para los Lag del dataframe “datos1” y “lag_features2” para los Lag del dataframe “datos2”.

Figura 39

Creaciones de Funciones para Lags y Rolling Window

```
#funcion para extraer lags
def get_n_lags(ts, n):
    lags = []
    for lag in range(1,n+1):
        lags.append(ts.shift(lag).rename(f'lag_{lag}'))
    return pd.concat(lags, axis=1)
```

```
#funcion que devuelva los rolling window
def get_window_mean(ts, n):
    windows = []
    for window in range(2,n+1):
        windows.append(ts.shift(1).rolling(window=window).mean().rename(f'mean_{window}'))
    return pd.concat(windows, axis=1)
```

```
lag_features1 = get_n_lags(datos1I.PAQUETES_SAP, 30) # de los últimos 30 días
lag_features2 = get_n_lags(datos2I.PAQUETES_SAP, 30) # de los últimos 30 días
```

Nota: Elaboración propia

A continuación, se procedió a concatenar los features de Lag y Rolling Windows, de los dataframe “datos1” y “datos2”, mediante la función “concat”, y a su vez, se incorporó el campo “PAQUETES_SAP” (target). Esto, con la finalidad de mejorar la precisión y el rendimiento de los modelos de Machine Learning, permitiendo la identificación de tendencias y patrones en los datos.

Como resultado de la concatenación de features, se obtuvo los dataframe “features1” y “features2”. El primero, contiene la unión de los datos de Lag y Rolling Windows del dataframe “datos1” y el segundo, datos del dataframe “datos2”.

Figura 40

Concatenación de los Features

```
features1 = pd.concat([lag_features1, window_features1], axis=1)
features1.head(3)
```

Fechas	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	lag_8	lag_9	lag_10	...	mean_21
2019-01-02	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
2019-01-03	2524.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
2019-01-04	2678.0	2524.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN

3 rows × 59 columns

Nota. Elaboración propia

Posteriormente, se procedió a adicionar el campo “PAQUETES_SAP” a los dataframe “features1” y “features2”, mediante la función “merge”, y se obtuvo los dataframe “features_target1” y “features_target2”.

Figura 41

Unión de la variable “PAQUETES_SAP”

```
features_target1 = features1.merge(datos1I.PAQUETES_SAP, left_index=True, right_index=True)# obtener el target: inner join
features_target2 = features2.merge(datos2I.PAQUETES_SAP, left_index=True, right_index=True)# obtener el target: inner join
```

Nota. Elaboración propia

De la aplicación de Lag y Rolling Windows de 30 días, se generaron datos nulos, por tanto, se procedió a eliminar las filas que contenían datos nulos; es decir, los datos de un mes. Este procedimiento, se realizó mediante la función “dropna”, la cual se utiliza para eliminar valores NaN (Not a Number) de filas y columnas en un dataframe.

Luego, se restablecieron los índices del dataframe “datos1” y “datos2” y se realizó la conversión del campo “fechas” a formato de fechas mediante la función “datetime” y mediante la función “day_name”, se extrajo el nombre del día de la semana de cada fecha.

A partir de esto, mediante el método de One Hot Encode, se procedió a convertir las variables categóricas en vectores binarios compatibles con algoritmos de Machine Learning. Para ello, se utilizó la función “get_dummies” de la librería pandas. Esta función, toma una columna de datos categóricos y crea una nueva columna para cada categoría en la columna original. Cada nueva columna contiene valores binarios (0 o 1) que indican si la fila pertenece a esa categoría o no.

Figura 42

Aplicación de One Hot Encode

```
#OneHotEncoding
dummies_day_name1 = pd.get_dummies(features_target1['day_name'], prefix="day_")
dummies_day_name1.head(2)
#----
dummies_day_name2 = pd.get_dummies(features_target2['day_name'], prefix="day_")
dummies_day_name2.head(2)
```

	day__Friday	day__Monday	day__Saturday	day__Sunday	day__Thursday	day__Tuesday	day__Wednesday
0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	0

Nota: Elaboración propia

Enfoque Regresión

En el presente enfoque, se llevó a cabo la agrupación del conjunto de datos en un dataframe. Las variables categóricas fueron agrupadas, mientras que las variables numéricas (“PAQ C/C”) fueron sumadas. Este proceso, es fundamental para la preparación de los datos y la construcción de modelos de pronóstico precisos y efectivos.

Luego, se procedió a convertir las variables categóricas en vectores binarios, ello mediante la función “get_dummies”. Posteriormente, se realizó la separación de los días, meses y años en columnas. Este proceso resultó en un dataframe con 37 columnas y 6,273 registros para cada producto.

Figura 43

Dataframe para los modelos con el Enfoque de Regresión

	PAQUETES_SAP	KMS	Otros	Regional Directas	Regional Distribuidoras	Regional E-commerce	Regional Interior	Regional Lima	Regional Mercados Especiales	Regional On Trade	...	March	May	November	October	September	anio_2019
0	71.0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	1
1	226.0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	1
2	2192.0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	1
3	10.0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	1
4	160.5	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6268	907.0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0
6269	648.0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0
6270	140.0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0
6271	3620.0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0
6272	3133.0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0

6273 rows × 37 columns

Nota: Elaboración propia

5.1.2.4 Modelamiento

En esta etapa, se construyeron 04 modelos de Machine Learning, desde los enfoques de Regresión y Forecasting, aplicando 08 técnicas de tratamiento de datos, ello, con el objetivo de identificar el modelo que mejor se ajuste a los datos de la demanda. A continuación, el detalle:

- ✓ Regresión lineal: Su aplicación radica en analizar la relación lineal entre las variables independientes y la variable dependiente (“PAQUETES_SAP”). Para este modelo, se utilizaron datos estandarizados y no estandarizados.
- ✓ Boosting LightGBM Regressor: El modelo de aprendizaje automático basado en árboles de decisión, se aplicó con el objetivo de analizar las relaciones no lineales entre las variables independientes y dependiente (“PAQUETES_SAP”). Para este modelo, se utilizaron datos estandarizados y no estandarizados.
- ✓ Modelo AutoArima con parámetros de modelo SARIMA: Su aplicación consiste en analizar las tendencias y patrones estacionales en los datos de la demanda. Para este modelo, se trabajará con datos con transformación logarítmica y sin transformación logarítmica.
- ✓ Modelo Prophet de Facebook: Modelo de código abierto utilizado en pronóstico de series temporales. La finalidad de su aplicación radica en experimentar el rendimiento del modelo. Del mismo modo que, el modelo SARIMA, se aplicará transformación logarítmica a los datos.

5.1.2.4.1 Regresión Lineal

Enfoque de Forecasting

La Regresión Lineal es un método de predicción que utiliza una variable dependiente (Y) y otras variables independientes (X) para encontrar una relación lineal que minimice el error de predicción. Para la construcción del modelo, se empleó la biblioteca "Sklearn" de Python. Luego, se procedió a separar las variables en dos conjuntos de datos.

X: que representa todas las variables independientes del modelo.
Y: que es la variable dependiente que se busca predecir.

Este procedimiento se aplicó para ambos productos. La Figura 44 muestra la separación de las variables de la bebida Evervess Ginger 1.5 litros de 06 unidades.

Figura 44

Separación de Variables

```
[47] X1 = features_target_with_day1[features_columns1]
     y1 = features_target_with_day1[target1]
```

Nota. Elaboración propia

A continuación, se procedió a dividir los datos en conjuntos de entrenamiento y de prueba. El conjunto de entrenamiento se utilizó para entrenar el modelo, es decir, para ajustar los parámetros del modelo a los datos, mientras que el conjunto de prueba se utilizó para evaluar el rendimiento del modelo en datos no vistos. Este proceso, se aplicó para las variables dependientes e independientes de ambos productos, considerando una fecha de corte. En este caso, la fecha de corte trazada fue el 01 de enero del 2023, manteniendo al campo “Fecha” como el índice en los dataframe. En la figura 45 se observa que, “X_train1” y “Y_train1”, corresponde al conjunto de datos de entrenamiento y “X_test1” y “Y_test1” al conjunto de datos de prueba de la bebida Evervess Ginger 1.5 litros de 06 unidades.

Figura 45

División de datos de entrenamiento y de prueba

```
[ ]  #Train and test: forecasting
     fecha_corte = "2023-01-01"# PARA EL ÚLTIMOS MES (PREDICCIÓN DIARIA)
     X_train1 = X1.reset_index("Fechas").query(f"Fechas < '{fecha_corte}'").set_index("Fechas")
     y_train1 = y1.reset_index("Fechas").query(f"Fechas < '{fecha_corte}'").set_index("Fechas")
     X_test1 = X1.reset_index("Fechas").query(f"Fechas >= '{fecha_corte}'").set_index("Fechas")
     y_test1 = y1.reset_index("Fechas").query(f"Fechas >= '{fecha_corte}'").set_index("Fechas")
```

Nota: Elaboración propia

Luego, se procedió a crear dos instancias de la clase “LinearRegression” de la librería Sklearn. La primera instancia “alumno1” refiere al modelo de datos que corresponden a la bebida Evervess Ginger 1.5 litros de 06 unidades, y la segunda instancia “alumno2”, corresponde al modelo de datos de la bebida Concordia de Piña de 03 litros de 04 unidades. Seguidamente, se ajustó los modelos de Regresión Lineal a los datos de entrenamiento. Este procedimiento se observa en la Figura 46.

Figura 46

Creación del modelo de Regresión lineal

```
#creando regresion lineal
alumno1=LinearRegression()
#creando regresion lineal
alumno2=LinearRegression()
```

```
alumno1.fit(X_train1, y_train1)
alumno2.fit(X_train2, y_train2)
```

```
▾ LinearRegression
LinearRegression()
```

Nota. Elaboración propia

Posteriormente, se realizó la predicción con las variables independientes de prueba de las bebidas Evervess Ginger 1.5 litros de 06 unidades (X_test1) y Concordia de Piña de 03 litros de 04 unidades (X_test2). Asimismo, se adicionó un campo “ForecastRL” al dataframe “Y_test1” y “Y_test2”, que contiene la variable dependiente de prueba de cada producto y a este campo se le asignó los valores de las predicciones. Este proceso se puede observar en la Figura 47.

Figura 47

Predicciones

```
y_RL_pred1 = alumno1.predict(X_test1)
y_RL_pred2 = alumno2.predict(X_test2)

y_test1["ForescastRL"] = y_RL_pred1
y_test2["ForescastRL"] = y_RL_pred2
```

Nota. Elaboración propia

Regresión Lineal: Datos con estandarización

En este proceso, se dividieron los datos en conjuntos de datos de entrenamiento y de prueba, y se crearon dos instancias de StandarScaler () para ambas bebidas. Esta función centra los datos en la media y los escala por la desviación estándar. Una vez creada las instancias por cada conjunto de datos de cada bebida, el modelo procede a aprender los parámetros de la desviación estándar y la media de cada conjunto de variables dependiente e independiente. La Figura 48, muestra el proceso de escalamiento de las variables y el aprendizaje de los parámetros de estandarización de ambos productos.

Figura 48

Aprendizaje de parámetros de estandarización por variables

```
# Escalamiento
scaler_X1 = StandardScaler()
scaler_y1 = StandardScaler()
scaler_X2 = StandardScaler()
scaler_y2 = StandardScaler()

# Aprendemos parametros de escalamiento
scaler_X1.fit(X1_train)
scaler_y1.fit(y1_train)

scaler_X2.fit(X2_train)
scaler_y2.fit(y2_train)
```

```
▾ StandardScaler
StandardScaler()
```

Nota: Elaboración propia

Una vez que el modelo aprendió los parámetros de desviación estándar y la media por cada conjunto de datos, se crearon nuevos dataframe que almacenen los valores estandarizados de las variables dependientes e independiente de los conjuntos de entrenamiento y prueba para ambos productos. Como consecuencia de este proceso, se perdió el formato de fechas, por ende, se procedió a reindexarlas. El proceso de transformación e indexación de las variables de la bebida la bebida Evervess Ginger 1.5 litros de 06 unidades se observa en la Figura 49.

Figura 49

Estandarización de Variables

```
# Transformar en base al estimador
X1_train_scaled = pd.DataFrame(scaler_X1.transform(X1_train), columns=X1_train.columns)
y1_train_scaled = pd.DataFrame(scaler_y1.transform(y1_train), columns=y1_train.columns)
X1_test_scaled = pd.DataFrame(scaler_X1.transform(X1_test), columns=X1_test.columns)
y1_test_scaled = pd.DataFrame(scaler_y1.transform(y1_test), columns=y1_test.columns)

X1_train_scaled = X1_train_scaled.set_index(X1_train.index)
y1_train_scaled = y1_train_scaled.set_index(y1_train.index)
X1_test_scaled=   X1_test_scaled.set_index(X1_test.index)
y1_test_scaled=   y1_test_scaled.set_index(y1_test.index)
```

Nota. Elaboración propia

Luego de estandarizar las variables, se crearon dos instancias de LinearRegression () para ambos productos. La instancia “alumno1N” contiene el modelo para los datos de la bebida Evervess Ginger 1.5 litros de 06 unidades, mientras que la instancia “alumno2N” contiene el modelo para los datos de la bebida Concordia de Piña de 03 litros de 04 unidades. A continuación, se entrenan los modelos con los datos de entrenamiento, usando las variables dependientes e independientes que corresponden a los conjuntos de datos para cada producto. El desarrollo de este proceso en la Figura 50.

Figura 50

Regresión Lineal con datos estandarizados

```
#creando regresion lineal
alumno1N=LinearRegression()
#creando regresion lineal
alumno2N=LinearRegression()
```

```
alumno1N.fit(X1_train_scaled,y1_train_scaled)
alumno2N.fit(X2_train_scaled,y2_train_scaled)
```

```
▾ LinearRegression
LinearRegression()
```

Nota. Elaboración propia

Posteriormente, se realizaron las predicciones de las variables independientes. Los resultados se encuentran en rango de -1 a 1, esto debido a la estandarización. Por ende, es necesario aplicar el proceso inverso de estandarización para obtener los valores reales y finalmente se reindexan las fechas en el dataframe. La Figura 51, muestra la predicción y des-escalamiento del elemento predicho de la bebida Evervess Ginger 1.5 litros de 06 unidades.

Figura 51

Predicción de Regresión Lineal con datos estandarizados

```
# Aplicamos modelo para predicciones
y_train_RL_pred1N = alumno1N.predict(X1_train_scaled)
y_test_RL_pred1N=alumno1N.predict(X1_test_scaled)

# De-escalamos el elemento predicho
y_train_pred_original1NRL = pd.DataFrame(scaler_y1.inverse_transform(y_train_RL_pred1N.reshape(-1,1)), columns=y1_train.columns)
#y_train_pred_original1N = scaler_y1.inverse_transform(y_train_lgb_pred1N.reshape(-1,1))
y_test_pred_original1NRL = pd.DataFrame(scaler_y1.inverse_transform(y_test_RL_pred1N.reshape(-1,1)),columns=y1_test.columns)
```

Nota. Elaboración propia

Enfoque de Regresión

En el desarrollo del modelo de Regresión Lineal, desde el enfoque de Regresión, se utilizaron las variables “Region”, “Atencion” y “Fecha”. Se dividieron las variables dependientes e independientes y se crearon conjuntos de datos de entrenamiento y de prueba. Para ello, se consideró un porcentaje de 80% de los datos para entrenamiento y 20% para prueba. Posteriormente se realizaron las predicciones de los datos. Este procedimiento, se aplicó para ambos productos. A continuación, el detalle:

Figura 52

Separación del conjunto de datos para entrenamiento y prueba

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Entrenar el modelo de regresión lineal
linear_model = LinearRegression()

# Hacer predicciones con el modelo en los datos de prueba
Y_pred_linear = linear_model.predict(X_test)
```

Nota. Elaboración propia

5.1.2.4.2 LightGBM Regressor

Enfoque de Forecasting

El modelo de LigthGBM Regressor se caracteriza por usar árboles de decisión y está optimizado para un sistema de alto rendimiento con sistemas distribuidos. El desarrollo, implicó crear dos instancias de LGBM Regressor para ambos productos, con la finalidad de entrenar las variables dependientes e independientes de los datos de entrenamiento. En la Figura 53 se observa que “lgb_model1” corresponde a la instancia de LGBM Regressor de la bebida Evervess Ginger 1.5 litros de 06 unidades, mientras que “lgb_model2” a la instancia de LGBM Regressor de la bebida Concordia de Piña de 03 litros de 04 unidades.

Figura 53

Modelo LightGBM Regressor

```
import lightgbm as lgb
lgb_model1 = lgb.LGBMRegressor()
lgb_model1.fit(X_train1, y_train1)
#-----
lgb_model2 = lgb.LGBMRegressor()
lgb_model2.fit(X_train2, y_train2)
```

```
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.001188 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 15059
[LightGBM] [Info] Number of data points in the train set: 1430, number of used features: 66
[LightGBM] [Info] Start training from score 3304.152447
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.001594 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 15059
[LightGBM] [Info] Number of data points in the train set: 1430, number of used features: 66
[LightGBM] [Info] Start training from score 1694.785664
▾ LGBMRegressor
LGBMRegressor()
```

Nota: Elaboración propia

Luego, se realizó la predicción con las variables independientes de prueba de la bebida Evervess Ginger 1.5 litros de 06 unidades (“X_test1”) y de Concordia de Piña de 03 litros de 04 unidades (“X_test2”). A continuación, la predicción de la variable independiente de la bebida Evervess Ginger 1.5 litros de 06 unidades.

Figura 54

Predicción con Modelo LigthGBM Regressor

```
[53] y_lgb_pred1 = lgb_model1.predict(X_test1)
```

Nota: Elaboración propia

LightGBM Regressor: Datos estandarizados

Para este modelo, se consideró el conjunto de datos del modelo de Regresión Lineal con datos estandarizados, y se crearon dos instancias de LGBMRegressor por cada producto. Luego, se procedió a entrenar el modelo con las variables del conjunto de datos de entrenamiento. Seguidamente, se realizaron las predicciones, las cuales varían entre -1 y 1, esto, debido a que los datos se encuentran estandarizados y finalmente, se transforman las predicciones a valores reales con el proceso de estandarización a la inversa. A continuación, la aplicación del modelo

LGBM Regressor para la bebida Evervess Ginger 1.5 litros de 06 unidades.

Figura 55

Modelo LigthGBM Regressor con data estandarizada

```
lgb_model1N = lgb.LGBMRegressor()
lgb_model1N.fit(X1_train_scaled, y1_train_scaled)

# Aplicamos modelo para predicciones
y_train_lgb_pred1N = lgb_model1N.predict(X1_train_scaled)
y_test_lgb_pred1N=lgb_model1N.predict(X1_test_scaled)

# De-escalamos el elemento predicho
y_train_pred_original1N = pd.DataFrame(scaler_y1.inverse_transform(y_train_lgb_pred1N.reshape(-1,1)), columns=y1_train.columns)
#y_train_pred_original1N = scaler_y1.inverse_transform(y_train_lgb_pred1N.reshape(-1,1))
y_test_pred_original1N = pd.DataFrame(scaler_y1.inverse_transform(y_test_lgb_pred1N.reshape(-1,1)),columns=y1_test.columns)
```

Nota: Elaboración propia

Enfoque de Regresión

En este enfoque, se utilizaron las variables “Region”, “Atencion” y “Fecha”, para luego separar las variables independientes. Asimismo, se crearon conjuntos de datos de entrenamiento y de prueba para ambos productos. Se consideró 80% de los datos para entrenamiento y 20% para prueba. Posteriormente, se realizaron las predicciones de la demanda.

5.1.2.4.3 Series Temporales

Para la implementación de series temporales, se utilizó el modelo ARIMA, con parámetros de SARIMA y modelo FB Prophet.

SARIMA

Para determinar si los datos presentan estacionalidad, se aplicó el Test de Dicky Fuller Aumentada, mediante la función “adfuller”, la cual, proporciona una tupla que contiene la estadística de prueba, el valor crítico para rechazar la hipótesis nula en diferentes niveles de significancia, el número de lagos utilizados en el modelo de autoregresión (AR) y los resultados de la prueba. Por tanto, si los datos presenten estacionalidad se utilizará el modelo SARIMA, caso contrario, ARIMA.

De la aplicación de la prueba de Dicky Fuller Aumentada, se obtuvo que los datos de la bebida Concordia de Piña de 03 litros de 04 unidades y Evervess Ginger de 1.5 litros de 06 unidades, presentan estacionalidad, por lo tanto, utilizó el modelo SARIMA. En la Figura 56 se observa los resultados de la bebida Evervess Ginger de 1.5 litros de 06 unidades.

Figura 56

Prueba de Dickey Fuller Aumentada

```
Prueba_Dickey_Fuller(datos1A['PAQUETES_SAP'],'PAQUETES_SAP')

Resultados de la prueba de Dickey-Fuller para columna: PAQUETES_SAP
Test Statistic                  -6.148691e+00
p-value                          7.655497e-08
No Lags Used                     1.300000e+01
Número de observaciones utilizadas   1.658000e+03
Critical Value (1%)             -3.434300e+00
Critical Value (5%)             -2.863285e+00
Critical Value (10%)            -2.567699e+00
dtype: float64
Conclusion:====>
Rechazar la hipótesis nula
Los datos son estacionarios
```

Nota. Elaboración propia

Luego, se procedió a dividir los datos de entrenamiento y los datos de prueba para ambos productos. Esta división está comprendida en 1,582 registros para entrenamiento y 90 registros de prueba. La figura 57 muestra la división de datos en entrenamiento y prueba de la bebida Evervess Ginger de 1.5 litros de 06 unidades.

Figura 57

División de datos en entrenamiento y prueba

```
train_data = datos1A[:len(datos1A)-90] # se esta poniendo 90
test_data = datos1A[len(datos1A)-90:]
test=test_data.copy()
```

Nota. Elaboración propia

A continuación, se importó la librería “pmdarima”, y utilizó la función “auto_arima”, la cual busca identificar los parámetros óptimos para un modelo ARIMA, es decir, la función “auto_arima”, ajusta el mejor modelo ARIMA a una serie temporal, según el criterio de información proporcionado. En este caso, el criterio utilizado fue el Criterio de Información de Akaike (AIC). Asimismo, esta función realiza una búsqueda (ya sea por pasos o en paralelo)

sobre posibles modelos y periodos estacionales dentro de las restricciones proporcionadas, y selecciona los parámetros que minimizan la métrica dada. En este proceso, se ajustó un total de 50 modelos diferentes, cada uno caracterizado por una combinación única de parámetros, con el propósito de encontrar las especificaciones de parámetros que ofrezcan el mejor ajuste a los datos observados, y como resultado se obtuvo 25 modelos para cada uno de los productos.

Asimismo, se obtuvo que, la combinación de parámetros óptima resultó en el AIC más bajo, expresado en ARIMA (4,1,0) (2,1,0) [12]. Esta elección de parámetros demostró ser óptima tanto para el producto Concordia de Piña de 03 litros de 04 unidades y Evervess Ginger de 1.5 litros de 06 unidades, en función de la eficacia de ajuste a los datos observados. A continuación, la combinación de hiperparámetros utilizados para obtener los 50 modelos de la bebida Evervess Ginger de 1.5 litros de 06 unidades.

Figura 58

Modelo SARIMA

```
[ ]  # Modelo Auto-Arima
     from pmdarima import auto_arima
```

```
[ ]  #modelo arima para el producto 1 evervest
     modelo_auto=auto_arima(train_data,start_p=0,d=1,start_q=0,
               max_p=4,max_d=2,max_q=4, start_P=0,
               D=1, start_Q=0, max_P=2,max_D=1,
               max_Q=2, m=12, seasonal=True,
               error_action='warn',trace=True,
               supress_warnings=True,stepwise=True,
               random_state=20,n_fits=50)
     print(modelo_auto)
```

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,1,0)[12]             : AIC=28882.403, Time=0.23 sec
 ARIMA(1,1,0)(1,1,0)[12]             : AIC=28286.757, Time=6.50 sec
 ARIMA(0,1,1)(0,1,1)[12]             : AIC=inf, Time=11.28 sec
 ARIMA(1,1,0)(0,1,0)[12]             : AIC=28638.792, Time=0.37 sec
 ARIMA(1,1,0)(2,1,0)[12]             : AIC=28120.878, Time=21.96 sec
 ARIMA(1,1,0)(2,1,1)[12]             : AIC=inf, Time=25.92 sec
 ARIMA(1,1,0)(1,1,1)[12]             : AIC=inf, Time=8.69 sec
 ARIMA(0,1,0)(2,1,0)[12]             : AIC=28391.913, Time=3.88 sec
 ARIMA(2,1,0)(2,1,0)[12]             : AIC=27978.078, Time=18.83 sec
 ARIMA(2,1,0)(1,1,0)[12]             : AIC=28162.413, Time=6.18 sec
 ARIMA(2,1,0)(2,1,1)[12]             : AIC=inf, Time=30.39 sec
 ARIMA(2,1,0)(1,1,1)[12]             : AIC=inf, Time=11.92 sec
 ARIMA(3,1,0)(2,1,0)[12]             : AIC=27934.560, Time=20.39 sec
 ARIMA(3,1,0)(1,1,0)[12]             : AIC=28151.279, Time=2.51 sec
 ARIMA(3,1,0)(2,1,1)[12]             : AIC=inf, Time=28.12 sec
 ARIMA(3,1,0)(1,1,1)[12]             : AIC=inf, Time=12.38 sec
 ARIMA(4,1,0)(2,1,0)[12]             : AIC=27881.148, Time=24.42 sec
 ARIMA(4,1,0)(1,1,0)[12]             : AIC=28053.997, Time=12.39 sec
 ARIMA(4,1,0)(2,1,1)[12]             : AIC=inf, Time=36.69 sec
 ARIMA(4,1,0)(1,1,1)[12]             : AIC=inf, Time=18.79 sec
 ARIMA(4,1,1)(2,1,0)[12]             : AIC=inf, Time=42.40 sec
 ARIMA(3,1,1)(2,1,0)[12]             : AIC=inf, Time=35.28 sec
 ARIMA(4,1,0)(2,1,0)[12] intercept   : AIC=27883.148, Time=26.45 sec

Best model:  ARIMA(4,1,0)(2,1,0)[12]
Total fit time: 406.020 seconds
 ARIMA(4,1,0)(2,1,0)[12]
```

Nota: Elaboración propia

A partir de los parámetros obtenidos, se implementó el modelo con la asignación de valores específicos al modelo SARIMA. Para ello, se suministró información contenida en el conjunto de datos denominado "Train data". Este procedimiento, se realizó para ambos productos. Luego de entrenar el modelo, se adicionó el campo "Arima_Predictions" al dataframe "test_data1" y "test_data2" y se almacenó las predicciones en el dataframe "arima_pred1" y "arima_pred2". Este procedimiento se realizó mediante la función "predict". La Figura 59, muestra el proceso de predicción de la bebida Evervess Ginger de 1.5 litros de 06 unidades.

Figura 59

Predicción con SARIMA

```
[ ] #prediciones del producto evervest
    arima_pred = arima_result.predict(start = len(train_data), end = len(datos1A)-1, typ="levels").rename("ARIMA Predictions")
    arima_pred

    2023-05-03    3713.756239
    2023-05-04    2731.606394
    2023-05-05    2889.752920
    2023-05-06    2603.689103
    2023-05-07    3044.211456
                      ...
    2023-07-27    1859.072650
    2023-07-28    2685.510676
    2023-07-29    2051.430485
    2023-07-30    2564.461081
    2023-07-31    2826.923572
    Freq: D, Name: ARIMA Predictions, Length: 90, dtype: float64
```

Nota: Elaboración propia

SARIMA: Datos con transformación logarítmica

La aplicación de transformación logarítmica a los datos radica en estabilizar la variabilidad en los datos. Esta transformación logarítmica no modifica las características estacionales de la serie temporal. Asimismo, implica la conversión de valores numéricos que originalmente podrían ser muy grandes en valores más pequeños. Esto tiene el efecto de reducir la escala de los datos, lo que facilita el trabajo con números más manejables. En este caso, se aplicó transformación logarítmica a toda la secuencia de la serie temporal, es decir, se convirtió todo el conjunto de datos conocido como "[PAQUETES SAP]" en logaritmos.

El siguiente proceso, continúa las mismas etapas del proceso de modelado y predicción de datos sin transformación logarítmica. Luego, se revirtió el proceso de transformación logarítmica a través de la función exponencial.

Figura 60

Aplicación de Transformación logarítmica a los datos

```
[ ]  datos1A['PAQUETES_SAP']=np.log(datos1A['PAQUETES_SAP'])
     datos1A.head()
```

FECHA_FACTURA	PAQUETES_SAP
2019-01-02	7.833600
2019-01-03	7.892826
2019-01-04	8.391857
2019-01-05	8.947546
2019-01-06	7.954723

Nota: Elaboración propia

FB Prophet

Para la implementación del modelo Prophet de Facebook, se procedió a preparar los datos en un formato adecuado, esto implicó la modificación de los nombres de las columnas de ambos productos. La columna "FECHA_FACTURA", se renombró como "ds", mientras que, la variable objetivo que corresponde a "PAQUETES_SAP" se denominó con el nombre de "y"

Figura 61

Renombre de Campos

```
[ ]  datos1C=datos1C.reset_index()
     datos2C=datos2C.reset_index()
```

```
[ ]  df_fb=datos1C.rename(columns={"FECHA_FACTURA":"ds", "PAQUETES_SAP":"y"} )
     df_fb.head()
```

	ds	y
0	2019-01-02	2524.0
1	2019-01-03	2678.0
2	2019-01-04	4411.0
3	2019-01-05	7689.0
4	2019-01-06	2849.0

Nota: Elaboración propia

Luego, se realizó la manipulación de los datos, dividiéndolos en conjuntos de datos de entrenamiento y de prueba. Para esta división, se consideró 90 registros para prueba y 1,582 para entrenamiento.

Figura 62

Entrenamiento de FB Prophet

```
train_data_pr = df_fb.iloc[:len(datos1C)-90]
test_data_pr = df_fb.iloc[len(datos1C)-90:]
```

Nota Elaboración propia

A continuación, se procedió a crear los modelos FB Prophet con la configuración de los hiper parámetros necesarios para llevar a cabo el proceso de entrenamiento. La elección de estos parámetros se realizó en función de la estacionalidad inherente de los datos y las características específicas del problema.

Figura 63

Creación de Modelo FB Prophet

```
from prophet import Prophet
```

```
#modelo producto Evervest
m = Prophet(growth='linear',
    changepoints=None,
    n_changepoints=25,
    changepoint_range=0.8,
    yearly_seasonality=True,
    weekly_seasonality='auto',
    daily_seasonality='auto',
    holidays=None,
    seasonality_mode='additive',
    seasonality_prior_scale=10.0,
    holidays_prior_scale=10.0,
    changepoint_prior_scale=0.05,
    mcmc_samples=0,
    interval_width=0.8,
    uncertainty_samples=1000)

# Hacemos el entrenamiento
m.fit(train_data_pr)
```

Nota: Elaboración propia

Una vez, configurado y entrenado el modelo, se procedió a realizar las predicciones de ambos productos. La obtención de predicciones resultó en un conjunto de datos extenso, entre ellos, las columnas "yhat_lower" y "yhat_upper", representan los límites inferior y superior del intervalo de predicción, respectivamente. Asimismo, el campo "Yhat", refiere a la predicción puntual de la variable dependiente (y). La Figura 64 muestra la instancia de la predicción de la bebida Evervess Ginger de 1.5 litros de 06 unidades.

Figura 64

Instancia Predicción

	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	multiplicative_terms_upper	yhat
1667	2023-07-27	3351.443097	1518.016195	5040.846047	3334.459413	3366.782860	0.0	3347.461757
1668	2023-07-28	3350.191402	1608.381102	5220.651818	3332.842709	3365.715329	0.0	3307.832575
1669	2023-07-29	3348.939708	1456.931822	4934.370958	3331.322512	3364.647799	0.0	3197.356272
1670	2023-07-30	3347.688013	1281.809561	4744.229446	3329.950429	3363.710154	0.0	2999.766595
1671	2023-07-31	3346.436318	1278.154244	4983.207290	3328.049253	3362.586922	0.0	3171.850421

Nota. Elaboración propia

FB Prophet: Datos con transformación logarítmica

El conjunto de datos de ambos productos, fueron sometidos a una transformación logarítmica. Los resultados derivados de esta estrategia se detallarán en la sección subsiguiente, donde se llevarán a cabo las evaluaciones del desempeño del modelo. Esta decisión de emplear datos transformados en logaritmo forma parte del proceso de análisis y modelado, y permitirá una evaluación más precisa y efectiva del rendimiento del modelo FB Prophet.

5.1.2.5 Evaluación del modelo

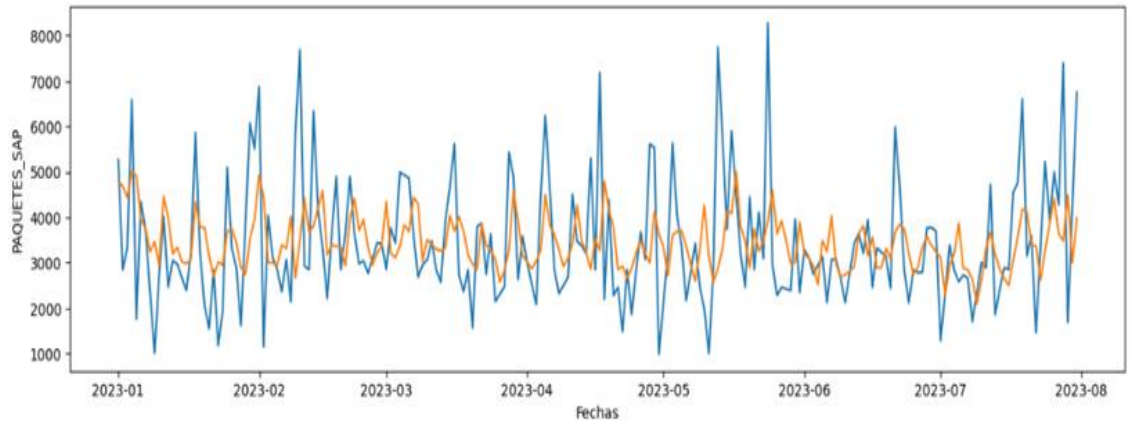
En esta etapa, se realizó la comparación de los modelos con datos estandarizados y sin estandarizar.

Para el modelo de Regresión Lineal y LightGBM Regressor sin estandarización, se obtuvo que los datos pronosticados se ajustaron a los datos reales, esto ocurre para los dos productos de estudio. Sin embargo, no se logró predecir los picos altos de los datos. Lo mismo ocurre para los datos estandarizados, se obtuvo que no mejoran al respecto al conjunto de datos

sin estandarizar. En las Figuras 65, 66, 67 y 68, las líneas de color azul representan los valores reales, mientras que las líneas de color naranja representan las predicciones.

Figura 65

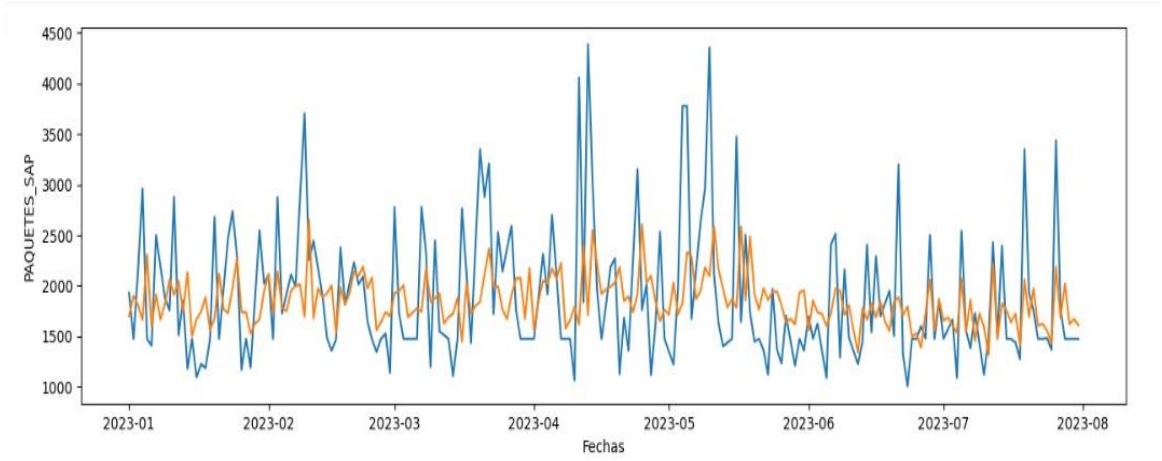
Regresión Lineal de la bebida Evervest Ginger



Nota. Elaboración propia

Figura 66

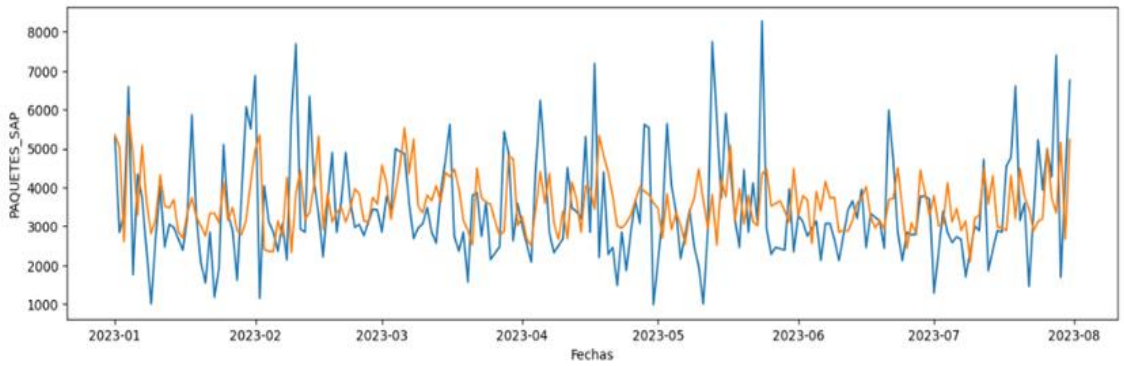
Regresión Lineal de la bebida Concordia de Piña



Nota. Elaboración propia

Figura 67

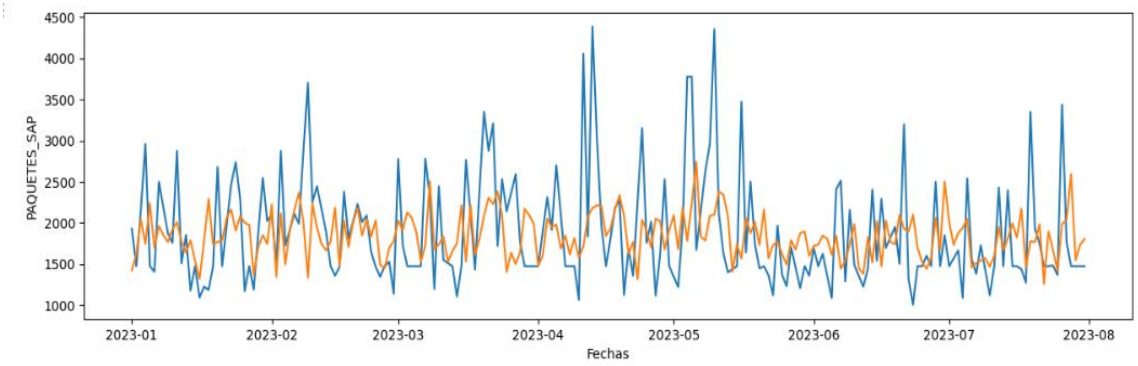
LGBM Regressor de la bebida Evervest Ginger



Nota. Elaboración propia

Figura 68

LGBM Regressor de la bebida Concordia de Piña

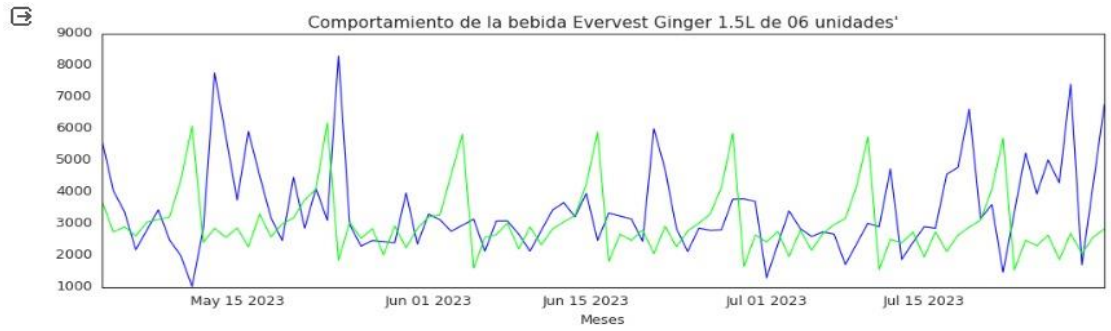


Nota. Elaboración propia

En el caso del modelo SARIMA, se obtuvo que los datos sin transformación logarítmica no tuvieron un buen ajuste, mientras que los datos con transformación logarítmica suelen mejorar, sin embargo, el modelo FB Prophet, obtiene mejores resultados. En la Figura 69 y 70, las líneas de color azul representan los valores reales, mientras que las líneas de color verde representan las predicciones.

Figura 69

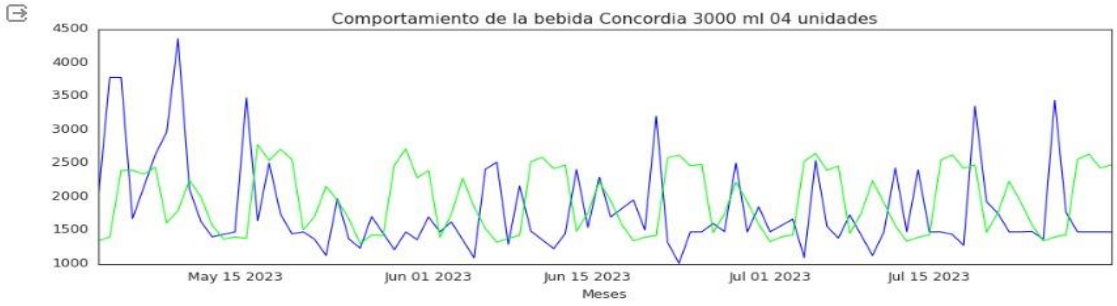
Modelo SARIMA de la bebida Evervest Ginger



Nota. Elaboración propia

Figura 70

Modelo SARIMA de la bebida Concordia de Piña

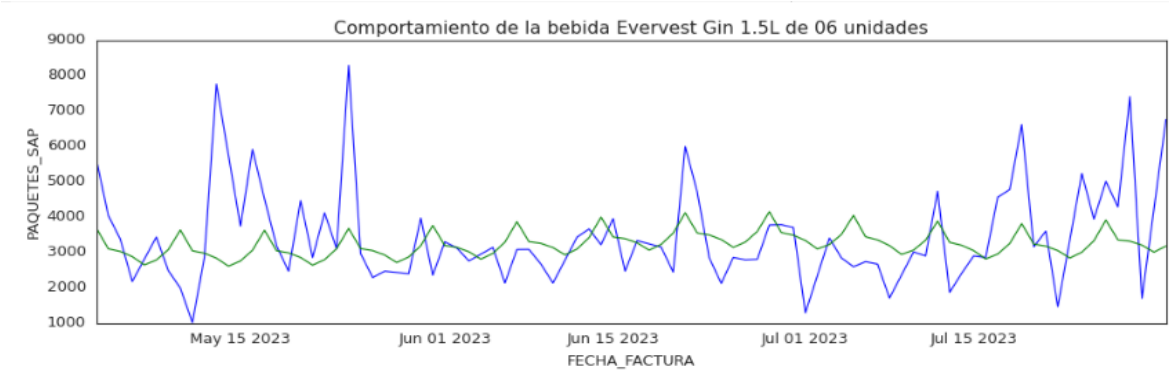


Nota. Elaboración propia

Para el modelo de FB Prophet, los datos pronosticados obtuvieron mejor ajuste respecto a los datos reales, este rendimiento representado en las métricas que se detallan en el siguiente apartado. En las Figuras 71 y 72, las líneas de color azul representan los valores reales, mientras que las líneas de color verde representan las predicciones.

Figura 71

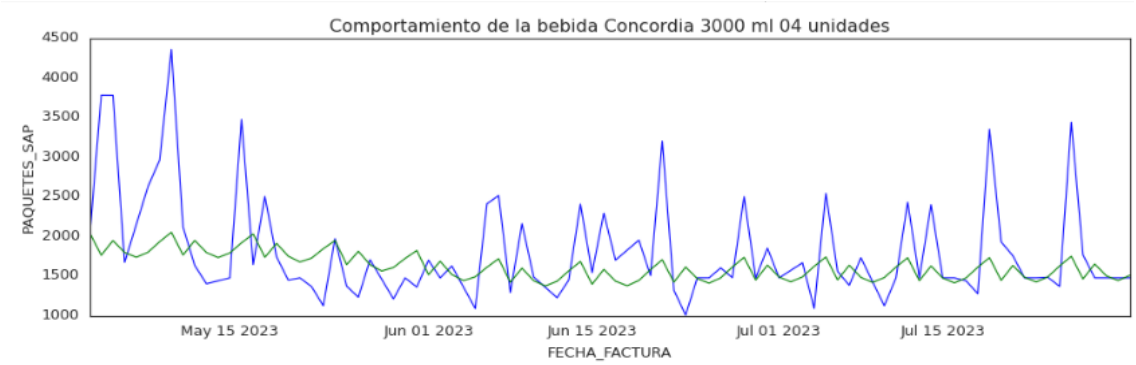
Modelo FB Prophet de la bebida Evervest Ginger



Nota. Elaboración propia

Figura 72

Modelo FB Prophet de la bebida Concordia de Piña



Nota. Elaboración propia

5.2 Medición de la solución

Para la medición de la solución propuesta, se utilizarán las métricas MAE (Mean Absolute Error), MAPE (Mean Absolute Pecertange Error), RSME (Root Mean Square Error), las cuales, se describen a continuación:

Mean Absolute Percentage Error (MAPE)

Makridakis, S Hibon, M. (2000), sostiene que, MAPE, “es una medida comúnmente utilizada en la evaluación de la precisión de los modelos de pronóstico. Proporciona una forma de medir el error relativo en términos de porcentaje, lo que permite comparar diferentes métodos

de pronóstico en diferentes escalas de datos"

José, J., Moreno, M., Pol, A. P., Abad, A. S., & Blasco, B. C. (2013), consideran que, “el MAPE tiene características importantes y deseables que incluyen confiabilidad, medida sin unidades, facilidad de interpretación, claridad de presentación, soporte de evaluación estadística y el uso de toda la información relacionada con el error”.

Asimismo, Lewis (1982), propone una tabla de interpretación de valores del MAPE, para datos industriales y comerciales, la cual se muestra a continuación:

Tabla 18

Interpretación de valores del MAPE

MAPE	Interpretación
<10	Pronósticos altamente precisos
Oct-20	Buen pronóstico
20-50	Pronósticos Razonables
> 50	Pronósticos Inexactos

Nota. Adaptado de Lewis (1982, p. 40)

Fórmula 7. Fórmula MAPE

$$MAPE = \frac{1}{n} * \sum_{t=1}^{n} \left| \frac{(yi - \hat{y}\imath)}{yi} * 100 \right|$$

Donde:

MAPE: Mean Absolute Percentage Error (Error Porcentual Absoluto Medio)

n: Número total de observaciones o datos.

yi: Valor real observado.

$\hat{y}\imath$: Valor pronosticado o predicho.

Root Mean Square Error (RMSE)

"El Error Cuadrático Medio de la Raíz (RMSE) es una medida que evalúa la precisión de un modelo de pronóstico o regresión al calcular la raíz cuadrada de la media de los errores al cuadrado entre las predicciones del modelo y los valores reales. Se utiliza comúnmente en estadísticas y análisis de datos para cuantificar la discrepancia entre las predicciones del modelo y los datos observados". (Montgomery, D.C., Peck, E.A., & Vining, G.G., 2012)

Fórmula 8. Fórmula RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}\imath)^2}$$

Donde:

Σ: Notación de suma, indicando que debes calcular la suma de los valores dentro de los paréntesis para todos los puntos de datos.
yi: El valor real u observado para el i-ésimo punto de datos.
$\hat{y}\imath$: El valor predicho para el i-ésimo punto de datos, generado por tu modelo.
n: El número total de puntos de datos.

Mean Absolute Error (MAE)

MAE, es una medida comúnmente utilizada en el análisis de pronóstico y regresión para evaluar la precisión de los modelos. Proporciona una medida robusta de la discrepancia promedio entre los valores pronosticados y los valores reales, sin verse afectada por valores atípicos o errores grandes" (Hyndman, R. J., & Athanasopoulos, G., 2018).

Fórmula 9. Fórmula MAE

$$MAE = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} | yi - \widehat{yi}|$$

Donde:

MAE: Mean Absolute Error (Error Absoluto Medio)
n: Número total de observaciones o datos.
yi: Valor real observado.
$\widehat{yi}$: Valor pronosticado o predicho.

5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo

La comparación de las métricas de los modelos de Regresión Lineal, LightGBM Regressor, SARIMA y FB Prophet, implica evaluar y comparar el rendimiento de estos modelos en términos de precisión, calidad y eficiencia utilizando métricas MAPE, RMSE y MAE. Esto es esencial para identificar los modelos que mejor se adapten a los datos y sean más adecuados para realizar predicciones y análisis de series temporales.

Para la comparación de los resultados se obtuvo los valores de las medias de los conjuntos de datos para ambos productos. La bebida Evervess Ginger 1.5 litros de 06 unidades registra una media de 3,326, mientras que la bebida Concordia de Piña 03 litros de 04 unidades, presenta una media de 1,805.1.

5.2.1.1. Regresión Lineal

De las métricas aplicadas al modelo de regresión lineal, se obtuvo lo siguiente:

- ✓ Mean Absolute Percentage Error (MAPE): Se obtuvo que para la bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, el MAPE asciende a 34,20%, y 24,41%, respectivamente. Estos valores, expresan un valor de pronóstico aceptable.
- ✓ Root Mean Square Error (RMSE): El RMSE para las bebidas Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, asciende a 1332,16 y

635,86 respectivamente. Estos valores indican que ambos productos se encuentran por debajo de los valores de la media, lo que sugiere que las predicciones del modelo no son extremadamente precisas, pero aun así proporcionan una estimación razonable de los valores reales.

- ✓ Mean Absolute Error (MAE): La bebida Evervess Ginger 1.5 litros de 06 unidades registra un MAE de 991,35, mientras que la bebida Concordia de Piña 03 litros de 04 unidades, de 466,96 respectivamente. Estos valores, se encuentran por debajo de la media de ambos productos.

5.2.1.2. LightGBM Regressor

Del cálculo de los indicadores del modelo de LightGBM Regressor se obtuvo lo siguiente:

- ✓ Mean Absolute Percentage Error (MAPE): La bebida Evervess Ginger 1.5 litros de 06 unidades, registra un MAPE de 38,01%, mientras que, la bebida Concordia de Piña 03 litros de 04 unidades, el 25,12%; ambos valores, se encuentran entre el rango de pronóstico aceptables. Sin embargo, el modelo de regresión lineal se posiciona como un mejor modelo para la predicción de la demanda.
- ✓ Root Mean Square Error (RMSE): La bebida Evervess Ginger de 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, registran un RMSE, de 1419,52 y 652,71, respectivamente, los cuales, se encuentran por debajo del conjunto de valores de la media.
- ✓ Mean Absolute Error (MAE). Se obtuvo que el MAE para la bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, corresponde a 1070,35 y 481,86, respectivamente; ambos valores, entre el conjunto de datos de la media.

5.2.1.3. SARIMA

El resultado de las métricas obtenidas de la construcción del modelo SARIMA se presentan a continuación:

- ✓ Mean Absolute Percentage Error (MAPE): Se obtuvo que el MAPE es de 38,18 % para la bebida Evervess Ginger 1.5 litros de 06 unidades y 36, 74 % para la bebida Concordia

de Piña 03 litros de 04 unidades, y ambos se encuentran en un rango aceptable. Estos valores, proporcionan una estimación razonable de los valores reales.

- ✓ Root Mean Square Error (RMSE): La bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, registran RMSE de 1945,28 y 900.11, inferiores a la media de los datos.
- ✓ Mean Absolute Error (MAE): La bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, registran MAE de 1,390.90 y 692.37, por debajo de la media de ambos productos.

5.2.1.4. FB Prophet

A continuación, el resultado de las métricas obtenidos del modelo FB Prophet:

- ✓ Mean Absolute Percentage Error (MAPE): Se obtuvo que, el MAPE asciende a 29.86% para la bebida Evervess Ginger 1.5 litros de 06 unidades y 19.42% para la bebida Concordia de Piña 03 litros de 04 unidades, ambos valores, se encuentran en un rango aceptable. Los valores obtenidos indican que, el modelo FB Prophet, presenta mejores valores de pronósticos aceptables en comparación al modelo SARIMA.
- ✓ Root Mean Square Error (RMSE): La bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, registran valores de 1945,28 y 900.11, inferiores a la media.
- ✓ Mean Absolute Error (MAE): La bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 3 litros de 04 unidades, consignan valores de 964.80 y 405.51**,** respectivamente, menores valores en comparación al modelo SARIMA.

Asimismo, se presenta un resumen de las métricas obtenidas a partir de la construcción de los modelos de Regresión Lineal, LightGBM Regressor, SARIMA y FB Prophet desde el enfoque de Forecasting y Regresión.

Tabla 19

Resumen de indicadores

Modelos/Métricas	Evervess Ginger 1.5 litros de 06 unidades			Concordia de Piña 3 litros de 04 unidades		
	MAPE	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**
Regresión Lineal - Forecasting	34.20	**1332.16**	991.35	24.41	635.86	466.95
Regresión Lineal -Estandarizada - Forecasting	34.33	1334.81	993.92	24.44	637.29	467.83
Regresión Lineal - Regresión	41.93	1628.50	1418.24	28.34	680.04	764.56
LightGBM Regressor - Forecasting	38.01	1419.52	1070.34	25.12	652.71	481.86
LightGBM Regressor – Estandarizada - Forecasting	38.50	1430.14	1085.40	24.80	643.87	476.62
LightGBM Regressor - Regresión	39.02	1582.52	1339.47	28.76	696.75	787.22
SARIMA	38.19	1945.28	1390.90	36.73	900.11	692.37
SARIMA logarítmico	42.66	2066.58	1509.13	37.18	858.97	651.31
FB Prophet	**29.86**	**1382.07**	**964.80**	**19.42**	**625.72**	**405.51**
FB Prophet logarítmico	**27.93**	1525.09	1031.81	20.13	624.73	410.92

Nota: *Elaboración propia*

De la evaluación de los valores obtenidos de las métricas MAPE, RMSE y MAE, se identificó que, el modelo FB Prophet, bajo el enfoque de Forecasting, proporcionó mejores resultados en comparación a los modelos Regresión Lineal, LightGBM Regressor y SARIMA. Se obtuvo que, para la bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, el valor del MAPE asciende a 29.86 % y 19.42%, respectivamente, lo cual nos indica que, estos valores se encuentran entre el rango de un pronóstico aceptable. Asimismo, se obtuvo que el RSME, asciende a 1382.07 y 625.72 y el MAE, a 964.80 y 405.51, cuyos valores son menores a la media de los productos Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades de la data de prueba. Por lo tanto, el modelo FB Prophet propuesto con datos sin transformación logarítmica, se ajusta a los datos de la demanda.

A partir de esto, se construyó el modelo FB Prophet para 03 productos adicionales que constituyen las ventas más representativas de la empresa CBC Peruana S.A.C: Pepsi Cola Pet de 0.75 litros de 12 unidades (Código de producto: BA003709), Triple Kola Pet de 0.5 litros de 15 unidades (Código de Producto: BA003736) y Seven Up Pet de 0.355 litros de 15 unidades (Código de Producto: BA003722). De los resultados obtenidos, el valor del MAPE para cada uno de los 03 productos, asciende a 49.93%, 46.24% y 38.39%, respectivamente, valores que se encuentran entre el rango de pronóstico razonable, lo cual nos indica que el modelo FB Prophet sigue siendo efectivo. Los resultados en el Anexo N°1.

El análisis proporcionado sugiere que el modelo FB Prophet es una herramienta útil para obtener la predicción de la demanda de los productos bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades. Asimismo, aplicable para establecer pronósticos de productos adicionales. Los valores del MAPE obtenidos, indican que las predicciones del modelo proporcionan una estimación razonable de los valores reales.

5.2.2 Simulación de solución. Aplicación de Software

De la implementación de los modelos, el modelo FB Prophet obtuvo mejores resultados, por lo tanto, este modelo se implementará para obtener las predicciones de 03, 06 y 12 meses del año 2024. Para ello, se implementó un módulo, que permite cargar un archivo input con las fechas a predecir, y se obtuvo como resultado las predicciones requeridas.

Figura 73

Módulo de Predicciones a futuro con FB Prophet

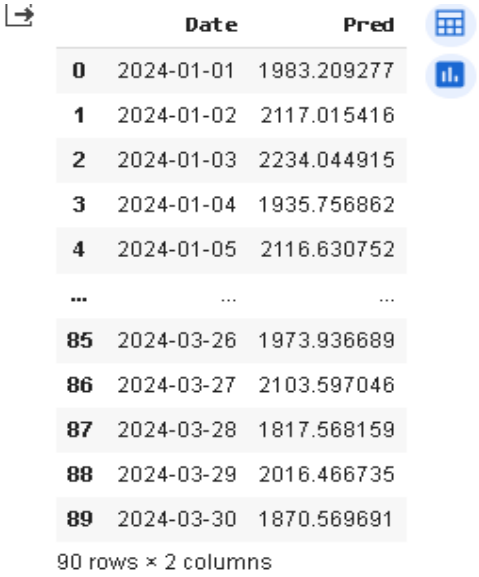
```
# import streamlit as st
from prophet import Prophet
from prophet.plot import plot_plotly, plot_components_plotly
import pickle
import pandas as pd
# Cargar los datos
# @st.cache
def load_data(filename, sheet_name):
    data = pd.read_excel(filename, sheet_name=sheet_name, header=None)
    data.columns=["ds"]
    return data

new_data = load_data("Ventas_fin.xlsx", sheet_name="PruebaAnio")


# Crear el modelo Prophet
# @st.cache
def loadall(filename):
    with open(filename, "rb") as f:
      model=pickle.load(f)
    return model
model = loadall('ModelProphet.pkl')



# Hacer predicciones

future = new_data
forecast = model.predict(future)

prophet_pred = pd.DataFrame({"Date" : forecast[-365:]['ds'], "Pred" : forecast[-365:]["yhat"]})
prophet_pred
```

Nota. Elaboración propia

Es preciso mencionar que, las predicciones que brinda este módulo se encuentran en frecuencia diaria, puesto que el modelo fue construido y entrenado en días. Para que el módulo funcione, es necesario tener un dataset con los periodos a predecir, como el caso de estudio solicita hacer predicciones de 03, 06 y 12 meses. Los resultados se mostrarán a continuación.

Figura 74

Resultado 3 meses



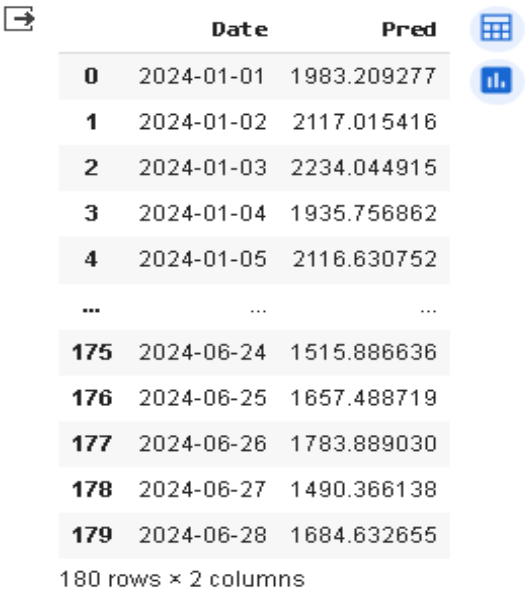
	Date	Pred
0	2024-01-01	1983.209277
1	2024-01-02	2117.015416
2	2024-01-03	2234.044915
3	2024-01-04	1935.756862
4	2024-01-05	2116.630752
...	...	...
85	2024-03-26	1973.936689
86	2024-03-27	2103.597046
87	2024-03-28	1817.568159
88	2024-03-29	2016.466735
89	2024-03-30	1870.569691

90 rows × 2 columns

Nota. Elaboración propia

Figura 75

Resultado a 6 meses



	Date	Pred
0	2024-01-01	1983.209277
1	2024-01-02	2117.015416
2	2024-01-03	2234.044915
3	2024-01-04	1935.756862
4	2024-01-05	2116.630752
...	...	...
175	2024-06-24	1515.886636
176	2024-06-25	1657.488719
177	2024-06-26	1783.889030
178	2024-06-27	1490.366138
179	2024-06-28	1684.632655

180 rows × 2 columns

Nota. Elaboración propia

Figura 76

Resultado de 12 meses

	Date	Pred
0	2024-01-01	1983.209277
1	2024-01-02	2117.015416
2	2024-01-03	2234.044915
3	2024-01-04	1935.756862
4	2024-01-05	2116.630752
...	...	...
360	2024-12-26	1976.506512
361	2024-12-27	2181.972831
362	2024-12-28	2032.279228
363	2024-12-29	1974.311365
364	2024-12-30	2043.921709

365 rows × 2 columns

Nota. Elaboración propia

CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

Conocer la demanda de un producto a futuro es fundamental para cualquier empresa, ya que permite tomar decisiones informadas sobre la producción, el inventario, la cadena de suministro y la estrategia de marketing. Al conocer la demanda, las empresas pueden anticiparse a las necesidades de los clientes y garantizar que los productos estén disponibles cuando los clientes los necesiten. Además, la demanda también puede afectar los costos de producción y los precios de venta, lo que puede tener un impacto significativo en la rentabilidad de la empresa.

De la realidad problemática, se identificó que, la empresa CBC Peruana S.A.C, registra altos costos de inventario, producto de errores en la planificación de la demanda de sus productos. Entre ellos, se identificó que, la demanda real de las bebidas Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, presentan mayor variabilidad respecto a la demanda proyectada. Como consecuencia, se genera sobre stock y en algunos casos, desatender los pedidos a los clientes. Por lo tanto, ante esta situación, se propuso como solución aplicar modelos de Machine Learning bajo los enfoques de Regresión y Forecasting, que permitan planificar la demanda de los productos que presentan mayor variabilidad.

Para el desarrollo e implementación de modelos de los Modelos de Machine Learning, se obtuvo información de las ventas históricas comprendida desde enero del 2019 a julio del 2023. A partir de esta información, se realizó la exploración de las variables y el análisis del comportamiento a través de diferentes periodos de tiempos para ambos productos. Del análisis, se obtuvo que las variables numéricas, muestran una correlación alta positiva entre ellas, las cuales descartamos para evitar redundancia, mientras que, del total de las variables categóricas, dos las variables muestran un comportamiento favorable frente a la variable objetivo.

Luego, se realizó el preprocesamiento de los datos con diferentes técnicas, entre ellas: interpolación, agrupamiento, eliminación de variables y tratamiento de outliers. Posteriormente, se construyeron cuatro modelos de Machine Learning: Regresión Lineal LGBM Regressor, SARIMA y FB Prophet, bajo los enfoques de Forecasting, Regresión, con datos estandarizados y sin estandarizar.

Es preciso señalar que, para la construcción de los modelos, se realizaron pruebas con diferentes frecuencias de datos, diaria y mensual, manteniendo un mejor rendimiento los datos con frecuencia diaria.

De la comparación de los resultados, se obtuvo que, el modelo FB Prophet, bajo el enfoque de Forecasting, logró un mejor ajuste en sus predicciones en comparación a los modelos de Regresión Lineal, LGBM Regressor y SARIMA. Para la bebida Evervess Ginger 1.5 litros de 06 unidades y Concordia de Piña 03 litros de 04 unidades, se obtuvo que el valor del MAPE asciende a 29.86 % y 19.42%, respectivamente, lo cual nos indica que, estos valores se encuentran entre el rango de un pronóstico aceptable del 20% al 30%. Asimismo, el RSME, asciende a 1382.07 y 625.72 y el MAE, a 964.80 y 405.51, cuyos valores son menores a la media de ambos productos. Asimismo, para evaluar la efectividad del modelo FB Prophet, se realizó la predicción de la demanda para los productos Pepsi Cola Pet de 0.75 litros de 12 unidades, Triple Kola Pet de 0.5 litros de 15 unidades y Seven Up Pet de 0.355 litros de 15 unidades, y se obtuvo que el modelo sigue siendo efectivo.

Finalmente, concluimos que, la implementación de técnicas de Machine Learning, mejorar la planificación de la demanda de la empresa CBC Peruana S.A.C.

6.2 Recomendaciones

A partir de la investigación realizada, se recomienda que, para futuras investigaciones de series temporales, es necesario realizar una investigación exhaustiva de los datos históricos. Esto implica recopilar información de 10 años históricos para identificar patrones estacionales, tendencias y ciclos que pueden ayudar a mejorar la precisión de las predicciones, además buscar si existe la posibilidad de encontrar variables exógenas que ayuden a los modelos a tener un mejor performance.

Asimismo, es recomendable contar con apoyo de expertos en la industria de bebidas, incluyendo profesionales de logística y producción. Estos expertos pueden proporcionar perspectivas únicas sobre los factores que afectan las ventas y ofrecer información valiosa sobre tendencias del mercado y comportamiento del consumidor.

Por otra parte, se recomienda explorar y experimentar con más modelos especializados en series temporales para la industria de bebidas, teniendo en cuenta las características únicas del sector. Esto puede incluir el análisis de series temporales multivariadas para diferentes tipos de productos y segmentos de mercado, así como la consideración de técnicas avanzadas

de aprendizaje profundo adaptadas a datos de ventas.

Finalmente se recomienda aplicar modelos más avanzados y complejos, como los que se encuentran en el campo del Deep Learning, como las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN), más técnicas de optimización las cuales sean capaces de capturar patrones complejos en datos temporales y que hayan demostrado ser eficaces en muchas aplicaciones de series temporales. Además, la técnica de ensamblaje de modelos, que combina las predicciones de varios modelos diferentes, a menudo produce resultados más precisos que un solo modelo.

Referencias Bibliográficas

Aamer, A., Eka Yani, L., & Alan Priyatna, I. (2020). Data analytics in the supply chain management: Review of Machine Learning applications in demand forecasting. Operations and Supply Chain Management: An International Journal, 14(1), 1-13.

Ballou, R. (2004). Logística, Administración de la Cadena de Suministros. (5ta Edición).

Bastarrica Lacalle, D. F. (2020). Predicción de series temporales mediante el método k-NN: explicabilidad y algoritmos de ensamblado.

Barrera, M. A. M. (Ed.). (s/f). Uso de la correlación De Spearman En Un Estudio De Intervención En Fisioterapia. Recuperado el 11 de noviembre de 2024, de http:Dialnet-UsoDeLaCorrelacionDeSpearmanEnUnEstudioDeIntervenc-5156978.pdf

Benos, L. (2021). Machine Learning in agriculture: a comprehensive updated review. Recuperado el 28 de Setiembre de 2023, de https://doi.org/10.3390/s21113758.

Beverage Market Size & Share Analysis - Growth Trends & Forecasts (2023 - 2028). Elaborador por Mordor Intelligence, Recuperado de https://www.mordorintelligence.com/industry-reports/beverages-market.

Brockwell, P.J. & Davis, R. A. (2016). Introduction to Time Series and Forecasting. 3rd Edition, Springer

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach. Springer.

Chatfield, C. (2000) Time Series Forecasting. Chapman and Hall, London.

Christopher, M. (2016). Logistics & supply chain management. Pearson UK.

Christophorus Beneditto, A., Darmawan, W., Bellatasya Unrica, N., Novita, H. (2021) Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET.

Chopra, S., & Meindl, P. (2015). Supply chain management: strategy, planning, and operation. Pearson.

Correa Loaiza, A. (2023). Análisis de modelos basados en Machine Learning para la predicción de la demanda de productos en la empresa Dyna & Cía. SA.

Cryer, J. D., & Chan, K. S. (2008). Time series analysis: with applications in R. Springer Science & Business Media.

Digital55. (2020, 5 de octubre). *Inteligencia Artificial, Machine Learning y Deep Learning*. Obtenido de https://digital55.com/blog/inteligencia-artificial-machine- learning-y-deep-learning/

Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci (2008). Introduction to Time Series Analysis and Forecasting.

EBAC Blog. (2023). Pronóstico de la demanda. Recuperado de https://ebac.pe/blog.

Elorza, H., & Medina Sandoval, J. C. (1999). Estadística para las ciencias sociales y del comportamiento. México: Oxford University

Emilio, N. (2022). What Types of Deep Learning Are There and What Are They For?. Recuperado de https://blog.bismart.com/en/tag/machine-learning.

Estela, M., & Javier, J. Reliability Models and Failure Detection Algorithms for Wind Turbines.

Fisher, M.L. (1997) What Is the Right Supply Chain for Your Product? A Simple Framework Can Help You Figure out the Answer. Harvard Business Review, 75, 105-116.

Gladys Choque Ulloa, Sandra Rosa Arroyo Paredes y Redy Rivas Idme. (2022, 6 de mayo). Modelos ARIMA, SARIMA y Método de Selección de Variables

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.st: "Time Series Analysis and Its Applications: With R Examples" de Shumway.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. Disponible en línea: https://otexts.com/fpp2/

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition.

Hyndman, R. J. (2018). Forecasting: principles and practice. Monash University, Australia. Disponible en línea: https://otexts.com/fpp2/

José, J., Moreno, M., Pol, A. P., Abad, A. S., & Blasco, B. C. (s/f). Using the R-MAPE index as a resistant measure of forecast accuracy. https://doi.org/10.7334/psicothema2013.23

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: un árbol de decisión que impulsa el gradiente altamente eficiente. Avances en sistemas de procesamiento de información neuronal, 30 (NIPS 2017), 3149-3157.

Khan, P. W., Byun, Y. C., Lee, S. J., & Park, N. (2020). Machine Learning based hybrid system for imputation and efficient energy demand forecasting. Energies, 13(11), 2681.

Kim, S. (2023). Innovating knowledge and information for a firm-level automobile demand forecast system: A Machine Learning perspective. Journal of Innovation & Knowledge, 8(2), 100355.

Lara Vizuete, S. A. (2022). Aplicación de técnicas de Machine Learning como método de validación para predecir la efectividad de un modelo estadístico de series de tiempo en la producción de fruta fresca en las diferentes provincias del Ecuador (Master's thesis, Universidad de Guayaquil-Facultad de Ciencias Matemáticas y Físicas-Carrera de Ingeniería Civil).

LASSO para Series Temporales (Parte 1). Ciencia de Datos Perú. Recuperado de https://datasciencepe.substack.com/p/modelos-arima-SARIMA-y-metodo-de

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). The bullwhip effect in supply chains. Sloan management review, 38(3), 93-102.

Lewis, C.D. (1982). Industrial and business forecasting methods. London: Butterworths

Lugon, A. (2023, febrero 23). Eliminar variables altamente correlacionadas del marco de datos en R (ejemplo). Estadisticool - La web de estaditica con Python y R. https://estadisticool.com/eliminar-variables-altamente-correlacionadas-del-marco-de-datos-en-r-ejemplo/

Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. International Journal of Forecasting, 16(4), 451-476.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: methods and applications (Vol. 3). John Wiley & Sons.

Martínez, W. R. (2020). Análisis de técnicas de Machine Learning aplicadas a la ciberseguridad informática para mejorar la detección de intrusiones y comportamientos anómalos en la Web.

McCarthy, J. (2007). What is artificial intelligence? Stanford University. Recuperado el 28 de agosto de 2023, de https://www- formal.stanford.edu/jmc/whatisai/whatisai.html

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2017). Estadística para negocios y economía (15ª ed.). Cengage Learning.

Miguéis, V. L., Pereira, A., Pereira, J., & Figueira, G. (2022). Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and Machine Learning. Journal of Cleaner Production, 359, 131852.

Mitchell, T. (1997). Machine Learning. McGraw Hill.

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

Parra, F. (2019). Estadística y Machine Learning con R, 48(3), 73-115.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., ... & Ziel, F. (2022). Forecasting: theory and practice. International Journal of Forecasting, 38(3), 705-871.

Producción Nacional. Elaborado por el Instituto Nacional de Estadística e Informática del Perú (abril del 2023). Recuperado de https://m.inei.gob.pe/media/MenuRecursivo/boletines/06-informe-tecnico-produccion-nacional-abr-2023.pdf

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine Learning-based approach: Global trends, research directions, and regulatory standpoints. Data Science and Management, 4(November), 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

Ramírez Morales, I. (2018). Estudio de aplicabilidad de técnicas de inteligencia artificial en el sector agropecuario.

Ranking de las principales empresas de bebidas a nivel mundial en función de su facturación en 2022 (29 de marzo del 2023). Portal Statista. Recuperado de https://es.statista.com/estadisticas/601277/mercado-global-de-bebidas-empresas-lideres-segun-ventas-netas/.

Reder, M. D. (2018). Reliability models and failure detection algorithms for wind turbines (Doctoral dissertation, Universidad de Zaragoza).

Rodríguez, IER (2022). Mejora del proceso de previsión de demanda en productos de consumo masivo en el mercado nacional de Agrosuper, mediante un enfoque de rediseño de procesos de negocios y minería de datos [PDF]. Recuperado de https://repositorio.uchile.cl/bitstream/handle/2250/185796/Mejora-del-proceso-de-pronostico-de-demanda-en-productos-de-consumo-masivo.pdf?sequence=1&isAllowed= y

Russell, S. J., & Norvig, P. (2010). Artificial intelligence: A modern approach

Shumway, R. H., & Stoffer, D. S. (2006). Time Series Analysis and Its Applications. Springer.

Shumway, R. H., & Stoffer, D. S. (2010). Time Series Analysis and Its Applications: With R Examples. Springer.

Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E. (2008). Designing and managing the supply chain: concepts, strategies, and case studies. McGraw-Hill.

Smith, J. (2005). Seasonal forecasting with SARIMA models. En Time Series Analysis for Forecasting (pp. 87-102). Springer.

Shumway, R. H., & Stoffer, D. S. (2017). Time series analysis and its applications: with R examples. Springer.

Trabajos de matemáticas. (octubre del 2023). Laminación-Análisis de ventana de modelos de series temporales. Recuperado de https://www.mathworks.com/help/econ/rolling-window-estimación-de-modelos-espaciales-de-estados.html

Tsay, R. S. (2010). Analysis of Financial Time Series (3rd ed.). Wiley.

Anexos

Anexo 01. Modelo FB Prophet

A partir de los resultados obtenidos, se procedió a construir el modelo FB Prophet, para 03 productos adicionales: Pepsi Cola Pet de 0.75 litros (750 mililitros) de 12 unidades, (Código de producto: BA003709), Triple Kola Pet de 0.5 litros de 15 unidades (Código de Producto: BA003736) y Seven Up Pet de 0.355 litros de 15 unidades (Código de Producto: BA003722).

En la etapa de recolección de información, se obtuvo data histórica desde el 01 enero del 2019 al 31 de julio del 2023. Luego, en la etapa de preprocesamiento, se realizó la exploración de los datos, y posterior verificación de campos nulos. Asimismo, en esta etapa, se procedió a renombrar las variables y modificó el tipo de dato para la columna “fecha”, y se obtuvo en formato de fecha. Para verificar si los datos del campo fechas se encuentran completos se aplicó la función de índice temporal (date_range). Para la verificación de campos nulos en el campo "Fechas" y "PAQUETES_SAP", se aplicó la técnica de Interpolación, mediante la función “interpolate”. Luego, con los datos completos, se procedió a construir el modelo FB Prophet, y para su evaluación, se obtuvieron las métricas MAPE, RMSE y MAE.

Los resultados, indican que, la bebida Pepsi Cola Pet de 0.75 litros de 12 unidades, Triple Kola Pet de 0.5 litros de 15 unidades y Seven Up Pet de 0.355 litros de 15 unidades, tienen un valor de MAPE, de 49.93, 46.24 y 38.39, respectivamente, lo cual indica que, se mantienen entre el rango de pronóstico razonable.

Figura 77

Modelo FB Prophet

▾ Prueba 3 datos adicionales con modelo *Prophet*

```
[330] import pandas as pd
      from prophet import Prophet
```

```
[331] datosP1 = pd.read_excel('Venta_Anexos.xlsx', sheet_name='BA003709')
      datosP2 = pd.read_excel('Venta_Anexos.xlsx', sheet_name='BA003736')
      datosP3 = pd.read_excel('Venta_Anexos.xlsx', sheet_name='BA003722')
```

```
[332] datosP1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1511 entries, 0 to 1510
Data columns (total 2 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Etiquetas de fila  1511 non-null   datetime64[ns]
 1   Suma de PAQ C/C    1511 non-null   int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 23.7 KB
```

```
[333] # Renombrar la columna 'A' a 'NuevaColumna'
      datos1A = datosP1.rename(columns={'Etiquetas de fila': 'FECHA_FACTURA','Suma de PAQ C/C': 'PAQUETES_SAP'})
      datos2A = datosP2.rename(columns={'Etiquetas de fila': 'FECHA_FACTURA','Suma de PAQ C/C': 'PAQUETES_SAP'})
      datos3A = datosP3.rename(columns={'Etiquetas de fila': 'FECHA_FACTURA','Suma de PAQ C/C': 'PAQUETES_SAP'})
```

```
[334] # Preparación del dato producto
      # ======================================================================
      datos1A['FECHA_FACTURA'] = pd.to_datetime(datos1A['FECHA_FACTURA'], format='%Y/%m/%d')
      datos1A = datos1A.set_index('FECHA_FACTURA')
      datos1A = datos1A.rename(columns={'x': 'y'})
      datos1A = datos1A.asfreq('d')
      datos1A = datos1A.sort_index()
```

```
[335] datos2A['FECHA_FACTURA'] = pd.to_datetime(datos2A['FECHA_FACTURA'], format='%Y/%m/%d')
      datos2A = datos2A.set_index('FECHA_FACTURA')
      datos2A = datos2A.rename(columns={'x': 'y'})
      datos2A = datos2A.asfreq('d')
      datos2A = datos2A.sort_index()
```

Nota. Elaboración propia

Figura 78

Métricas MAPE, RMSE, MAE

```
evaluacion_metrica(test_data_pr1["y"],test_data_pr1["Prophet_Predictions"])

Evaluation metric results:-
MAE is : 1894.4557981322155
RMSE is : 2266.717686797416
MAPE is : 49.93364240529179
```

```
[355] evaluacion_metrica(test_data_pr2["y"],test_data_pr2["Prophet_Predictions"])

      Evaluation metric results:-
      MAE is : 1676.6811691607534
      RMSE is : 2088.3360856022964
      MAPE is : 46.24256220600682
```

```
[356] evaluacion_metrica(test_data_pr3["y"],test_data_pr3["Prophet_Predictions"])

      Evaluation metric results:-
      MAE is : 987.7868995448968
      RMSE is : 1152.503785930625
      MAPE is : 38.39027451198321
```

Nota. Elaboración propia