Aplicación de la ciencia de datos en la toma de decisiones empresariales en el Sector Retail

Victor Andrés Lasso Vivas

Asesor

Sixto Enrique Campana

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnologías e Ingenierías ECBTI

Maestría en Gestión de Tecnología de Información

2024

Agradecimientos

A Dios, en quien está puesta mi esperanza, sin su guía y protección, el desarrollo de este proyecto no habría sido posible.

A mis padres, cuyo amor continuo, paciencia y apoyo incondicional me han sido invaluables a lo largo de este viaje. Su ejemplo de fe y nobleza es una fuente constante de inspiración.

A mis hermanos, en especial a mi hermano mayor, cuyos consejos y recuerdos me han motivado y han sido fundamentales en este proceso. Este logro es el reflejo del apoyo y el amor que me han brindado.

Resumen

En un entorno altamente competitivo, cambiante y saturado, los negocios minoristas del país se enfrentan al desafío de obtener ventajas y tomar medidas para evitar pérdidas o incluso luchar contra la desaparición. Al aprovechar las capacidades de la ciencia de datos y las nuevas herramientas de análisis de datos que contribuyen a la toma de decisiones informadas, una pequeña y mediana empresa (PYME) puede tomar decisiones encaminadas al aumento de su eficiencia y rentabilidad a partir del aprovechamiento de los datos generados en sus operaciones diarias. Sin embargo, las PYMEs se enfrentan a desafíos significativos debido al crecimiento en el volumen de datos y la diversidad de fuentes de información, lo que dificulta la comprensión y el uso efectivo de los recursos de datos disponibles. Como resultado, muchos de estos negocios continúan tomando decisiones basadas en la experiencia y la subjetividad de sus administradores. Por tanto, ésta investigación aporta diferentes soluciones para que las pymes superen estos desafíos para la toma de decisiones. Específicamente, en este estudio se abordan los siguientes aspectos: Se revisan los elementos metodológicos, técnicos y tecnológicos que definen la ciencia de datos, con el fin de identificar las características ideales de un entorno de datos propicio para la adopción y el desarrollo de la ciencia de datos. Además, se discute acerca de las ventajas y oportunidades que la aplicación de la ciencia de datos puede brindar al sector minorista, en cuanto, a la toma de decisiones estratégicas, examinando las tendencias actuales en su implementación y evaluando las tácticas, tecnologías y resultados obtenidos en estas aplicaciones. Por otro lado, se propone un método para el uso de la ciencia de datos que emplea la metodología CRISP-DM como marco de análisis y desarrollo de modelos para guiar el proceso de toma de decisiones. En definitiva, la solución se evalúa mediante la aplicación en un

caso de estudio, es decir, en un comercio minorista, demostrando su potencial para su implementación práctica dentro de la investigación.

Palabras clave**:** Data Science, Toma de Decisiones, Big Data, Comercio, Retail, Inteligencia artificial.

Abstract

In a highly competitive, changing and saturated environment, the country's retail businesses are faced with the challenge of gaining advantages and taking measures to avoid losses or even fight against disappearance. By leveraging the capabilities of data science and new data analysis tools that contribute to informed decision making, a small and medium-sized enterprise (SME) can make decisions aimed at increasing its efficiency and profitability by leveraging the data generated in its daily operations. However, SMEs face significant challenges due to the growth in the volume of data and the diversity of information sources, which makes it difficult to understand and effectively use the available data resources. As a result, many of these businesses continue to make decisions based on the experience and subjectivity of their managers. Therefore, this research provides different solutions for SMEs to overcome these decision-making challenges. Specifically, this study addresses the following aspects: The methodological, technical and technological elements that define data science are reviewed, in order to identify the ideal characteristics of a data environment conducive to the adoption and development of data science. In addition, the advantages and opportunities that the application of data science can provide to the retail sector are discussed, in terms of strategic decision-making, examining current trends in its implementation and evaluating the tactics, technologies and results obtained in these applications. On the other hand, a method for the use of data science is proposed that employs the CRISP-DM methodology as a framework for analysis and development of models to guide the decision-making process. Ultimately, the solution is evaluated through application in a case study, that is, in a retail store, demonstrating its potential for practical implementation within the research.

***Keywords*:** Data Science, Decision Making, Big Data, Commerce, Retail, Artificial Intelligence.

Tabla de Contenido

Lista de Figuras

Introducción

En el contexto empresarial los términos “minorista” y “Retail” se utilizan comúnmente para referirse a la misma actividad comercial. Según la definición técnica del comercio minorista, se refiere a la reventa (compra y venta sin transformación) de mercancías o productos, destinados para consumo o uso personal o doméstico (consumidor final) (DANE, 2019). El comercio minorista no sólo lo integra el canal de ventas tradicional conformado por tiendas de barrio y autoservicios pequeños; sino que abarca variadas líneas de mercancías relacionadas con alimentos (víveres en general), medicamentos, bebidas (alcohólicas y no alcohólicas), ropa, tecnología, elementos deportivos, entre otras actividades económicas. (Martínez M.A. 2022).

El sector minorista en Colombia ha ido en aumento en los últimos años, “en la región pacifica del país el crecimiento ha sido del 10% en los últimos 10 años” (Cruz C, 2017) y desempeña un papel fundamental en la economía del país reflejado en su participación en el PIB del 16%, superando a la industria cuya partición es del 11% y el sector inmobiliario del 8,43% (DANE, 2024). Esta situación se atribuye, en parte, a la arraigada cultura de los colombianos de realizar sus compras en las tiendas de barrio, al bajo nivel de ingresos y la dispersión de la población urbana. (Ríos et al., 2004).

Aunque los negocios minoristas se han consolidado como una opción conveniente y accesible para los consumidores en el país, enfrentan desafíos significativos, como las altas tasas de desempleo, que impulsan el crecimiento del comercio informal, y las elevadas cargas tributarias. Además, los minoristas deben lidiar con un entorno de alta competencia y problemas en el suministro de productos, lo que conlleva a la necesidad de tomar decisiones estratégicas con un mínimo margen de error y encontrar soluciones óptimas adaptadas a las características

particulares de cada negocio. En este contexto, surge el interés por aplicar disciplinas como la ciencia de datos en el proceso de toma de decisiones.

La Ciencia de Datos es un área interdisciplinaria que tiene como propósito transformar datos en valor para poder reportar, diagnosticar, predecir y también recomendar soluciones o mejoras en productos, servicios y/o procesos (Arriagada Benítez, M., 2020). Al revisar las fases del ciclo de vida en proyectos de ciencia de datos se puede ver que la mayoría de los autores terminan en una fase general normalmente denominada “uso de los datos”, la cual no se enfoca exclusivamente en la aplicación de los resultados para la toma de decisiones. Es aquí, donde metodologías como CRISP-DM cobran relevancia, ya que incluyen la evaluación del modelo y su impacto en la toma de decisiones dentro de una de sus etapas.

De esta manera, considerando el ciclo de vida de la ciencia de datos y apoyado en la metodología CRIPS-DM; se propone un modelo de analítica de datos para la para la toma de decisiones en el sector minorista. En este sentido, la metodología CRISP-DM abarcaría desde la comprensión de los objetivos empresariales hasta la evaluación y aplicación de los resultados, para luego utilizar estos resultados en la toma de decisiones en el ámbito minorista.

Descripción del Problema

Planteamiento del Problema

El entorno dentro del que se mueve el comercio minorista involucra varios desafíos como las altas tasas de desempleo (alrededor del 11%) (DANE, 2023) que conducen al crecimiento del comercio informal. Además, se enfrenta a elevadas cargas tributarias, (Colombia paga una de las tasas de tributación más altas de Latinoamérica con una tasa de crecimiento alrededor del 2,5%) (Ríos et al., 2004). Sin embargo, ante estas dificultades, surgen oportunidades para fortalecer la administración del negocio como posibilidades de acceso a crédito; plazos especiales para el pago de impuestos; y amnistías en los reportes ante centrales de riesgo (Fenalco, 2021).

Existen otros retos, que enfrentan los comercios minoristas diariamente, que parecen no tener una solución tan clara, los cuales se sintetizan en dos aspectos: una alta competencia y problemas con el suministro de productos. (Fenalco, 2021). Con el crecimiento de las ventas en línea, se observan cambios en el comportamiento de los consumidores, el aumento de las ventas al por menor, la alta competencia, problemas con el suministro de productos y el rápido aumento de los datos y fuentes de información provocado por las nuevas Tecnologías de la Información y la Comunicación (TIC).

Los minoristas se han enfrentado a una presión adicional para adoptar nuevas metodologías a la hora de tomar decisiones. (Aversa, J., Hernandez, T., & Doherty, S. 2021). En un contexto, donde la competencia es intensa y las condiciones del mercado pueden cambiar rápidamente, la capacidad de tomar decisiones informadas y ágiles se convierte en un factor clave para la supervivencia y el crecimiento del negocio. Para enfrentar estos retos, los dueños y gerentes de comercios minoristas deben adoptar un enfoque estratégico en la toma de decisiones que no sólo ofrezca certeza y confianza, sino que también minimice el margen de error. Esto

implica, implementar un proceso de toma de decisiones, basado en datos, que permitan evaluar de manera objetiva las opciones disponibles, y resulte clave para: “buscar alternativas eficaces para obtener liquidez y, así, adquirir una adecuada estabilidad financiera y dirigir las acciones hacia el crecimiento sostenible.” (Abad-Segura, E. et al., 2022)

Si venimos al contexto colombiano, en la mayoría de estos negocios aún se percibe una alta tendencia a utilizar la intuición y la experiencia del tomador de decisiones como herramientas fundamentales para resolver problemas (Cabeza de Vergara, L., & Muñoz, A. E., 2010). En la mayoría de los casos, las decisiones se basan únicamente en intuiciones u opiniones, careciendo de una metodología adecuada o de conocimientos sólidos que impulsen el crecimiento del negocio debido a que los directivos tienen dificultades en el conocimiento de los recursos de datos y cómo utilizarlos en sus negocios para tomar decisiones, especialmente en las PYMES (pequeñas y medianas empresas) (Henao Rosero, A., & Power, D. J., 2017).

A medida que estos negocios crecen, enfrentan desafíos relacionados con el uso de datos. La diversidad de proveedores en líneas de productos como abarrotes y frutas, añade una mayor complejidad a la gestión. Por lo tanto, es fundamental dar sentido a la gran cantidad de datos generados y emplear herramientas que faciliten la toma de decisiones, asegurando que estén alineadas con la estrategia y los objetivos del negocio para obtener soluciones relevantes y útiles.

Justificación

Las decisiones en el comercio minorista son cruciales, ya que pueden tener un impacto significativo en el éxito del negocio. Algunas decisiones son dinámicas, como la modificación de precios, la selección de productos a la venta o la planificación de campañas publicitarias. Sin embargo, también existen decisiones de alto riesgo, como la apertura de nuevas tiendas, la ampliación o renovación de establecimientos, y las adquisiciones, las cuales pueden afectar profundamente la rentabilidad y sostenibilidad del negocio, si no se toman con la debida consideración y análisis.

Es por esto, por lo que se hace necesaria la toma de decisiones en base a datos que ofrezcan un margen mínimo de error y resulten claves para: “buscar alternativas eficaces para obtener liquidez y, así, adquirir una adecuada estabilidad financiera y dirigir las acciones hacia el crecimiento sostenible.” (Abad-Segura, E. et al., 2022). En tal sentido, tomar decisiones en base a la extracción de información con herramientas informáticas y algoritmos, constituyen una forma de basarse en elementos cuantitativos que: “a partir de modelos de decisión, permiten obtener información que facilita las decisiones que se adaptan a los cambios del mercado, previenen acontecimientos futuros como operativos, y dan herramientas para la gestión diaria.” (Holsapple et al., 2014).

Además, permite aumentar la efectividad organizacional inteligente, analizar la toma de decisiones en las organizaciones y aumentar el rendimiento de la cadena de suministro existentes (Gawankar et al., 2020), así como consultar de forma óptima sus datos (Gallego-Gómez & De-Pablos-Heredero., 2017), mejorando las operaciones y maximizando la rentabilidad.

En los últimos años, se ha investigado sobre la implementación de métodos y prácticas basadas en tendencias tecnológicas para emplear de manera óptima los datos en la toma de

decisiones. Se han utilizado herramientas de Big Data (Gutiérrez Barrera, 2018), análisis predictivo (Chávez J. & Saucedo N., 2018), Analítica (Medina, E. J. 2021), e inteligencia artificial (Chávez E. et al., 2018) sobre todo en entornos minoristas, donde se ha visto: “la necesidad de conocer a los clientes y ofrecer productos de acuerdo con sus necesidades.” (Chiang, L. L. & Yang, C. S., 2018).

Aunque estas técnicas son fundamentales para el diseño y evaluación de algoritmos aún se ve: " la necesidad de realizar más investigaciones para acercar los resultados del proceso de ciencia de datos a las necesidades de los tomadores de decisiones comerciales." (Coussement K., & Benoit, F., 2021) ya que se ha dado mayor énfasis e interés a la práctica de la ciencia de datos en sí misma, descuidando la interpretabilidad y utilidad de los resultados para el negocio.

De ahí, la importancia de proponer un método de uso de datos descrito a través de un esquema secuencial que inicie con la comprensión del negocio, descubrimiento de conocimiento por medio de los datos, hasta la toma de decisiones veloces, basadas en un análisis cuantitativo, que permitirá la generación de información útil, favoreciendo los procesos de toma de decisiones, permitiendo sintetizar la información disponible. Con la finalidad de evaluar el entorno cambiante y las alternativas estratégicas; para planificar los recursos disponibles y emitir opiniones técnicas que faciliten el proceso de toma de decisiones. (Abad-Segura, E. et al., 2022).

Objetivos

Objetivo General

Desarrollar un modelo del proceso de analítica de datos basado en la metodología CRISP-DM y apoyado en ciencia de datos que permita guiar y optimizar el proceso de toma de decisiones en el sector Retail

Objetivos Específicos

Analizar las tendencias actuales en la aplicación de la ciencia de datos para la toma de decisiones en distintas áreas incluyendo el sector minorista, analizando los enfoques, tecnologías y resultados de estas aplicaciones.

Identificar los principales desafíos que enfrentan los negocios del sector Retail al implementar modelos de ciencia de datos.

Demostrar la importancia del uso de la ciencia de datos en la toma de decisiones en el sector Retail.

Explorar la aplicación de diferentes modelos y técnicas de ciencia de datos en la predicción de ventas y en la toma de decisiones estratégicas en el comercio minorista.

Proponer un modelo para analizar datos y apoyar la toma de decisiones en el sector Retail, aplicado en un estudio de caso del sector minorista del ecuador.

Aplicación de Ciencia de Datos en la Toma de Decisiones Empresariales en el Sector RETAIL

Estrategias y Técnicas de Aplicación de la Ciencia de Datos para la Toma de Decisiones

La ciencia de datos encuentra aplicación en una variedad infinita de situaciones donde la toma de decisiones es crucial, facilitando así decisiones informadas y mejorando la eficiencia en múltiples sectores. Por ejemplo, en entornos militares para fortalecer el rigor analítico y responder de manera más efectiva a las preguntas de inteligencia militar (Heilman, E., et al., 2019); en la salud para proporcionar planes médicos personalizados y para la predicción y el diagnóstico de enfermedades como el Alzheimer (Wu, C., & Xiao, L., 2021; Harper, M., et al., 2019), en la educación para prever la retención estudiantil en diferentes grupos demográficos de estudiantes (Li, C., et al., 2019).

A su vez, en decisiones comunitarias al facilitar información para abordar diversas necesidades locales y promover el involucramiento comunitario en las decisiones; (Chowdhury, M. T. A., & Sharma, N., 2021) presentando resultados de análisis de datos comprensibles y útiles para los responsables de tomar decisiones estratégicas, (Coussement K., & Benoit, F., 2021) También, en entornos gubernamentales, para mejorar los servicios públicos facilitando la identificación de patrones y tendencias, que pueden informar políticas y acciones más efectivas en diversos contextos sociales. (Paolotti, D., & Tizzoni, M., 2018).

Por último, en el sector primario en tareas como el análisis de viabilidad y productividad de un producto, control de salud de cultivos, control de plagas y control de calidad. Brindando información que permite al agricultor tomar decisiones rápidas basadas en el estado actual de la parcela. (Mompó Serrano, A., 2022). Las aplicaciones de la ciencia de datos en la toma de decisiones son variadas e incluyen el análisis de información, la realización de predicciones y la

recomendación de acciones basadas en datos. El uso de la ciencia de datos en la toma de decisiones se puede enfocar de tres formas: como guía, apoyo y automatización en la toma de decisiones.

Como Guía en la Toma de Decisiones

Llegando a este punto, la ciencia de datos sirve como un marco o una referencia para abordar los problemas, proporcionando un conjunto de herramientas conceptuales y modelos (descriptivos, predictivos y prescriptivos) que pueden aplicarse para analizar situaciones, evaluar las opciones y tomar decisiones. Aunque los problemas específicos pueden variar, el proceso de razonamiento es similar, ya que la esencia de la ciencia de datos, radica en el método científico que es: formular hipótesis y validar conclusiones. En otras palabras, trabajar con datos de manera científica. En este contexto, hablamos de aplicaciones centradas en el empirismo, que reacciona ante los datos a través de experimentos y preguntas, centrándose en el método empírico más que en las herramientas específicas.

Respecto a ello, Li, C., et al., (2019) utilizaron la ciencia de datos para formular y validar hipótesis sobre los factores que influyen en la retención estudiantil. Plantearon las hipótesis de que el rendimiento académico y la situación financiera de los estudiantes, están estrechamente relacionadas con su retención en la universidad, emplearon modelos de regresión Lineal y Logística, árboles de decisión y métodos de aprendizaje automático (kNN) para estudiar la retención estudiantil, y finalmente a partir de estos métodos, identificaron las variables más importantes para predecir la retención estudiantil y lograron una alta precisión en la predicción de la retención de estudiantes en la universidad.

Otro ejemplo de este uso, se puede ver en la investigación de Heilman, E., et al. (2019), donde se propone el uso de la ciencia de datos en los niveles superiores del Ejército, para

responder preguntas de inteligencia. En esta investigación, se sugiere aplicar la ciencia de datos de manera estructurada y científica en el análisis de inteligencia militar, guiando la toma de decisiones, formulando hipótesis y validando conclusiones a través de la metodología CRISP-DM. Buscando aumentar la confianza de las estimaciones a un nivel superior y proporcionar productos de inteligencia mejorados.

En el contexto empresarial, Bocangel, J. L., et al. (2020), usan la ciencia de datos para responder preguntas como: ¿Es posible predecir si un cliente Freemium se convertirá en Premium? La hipótesis que plantean sugiere que utilizando un modelo de árbol de decisión se puede predecir la conversión de cuentas Freemium a Premium en una empresa peruana, basándose en el comportamiento de diversas variables. Los resultados revelaron que el modelo de árbol de decisión efectivamente permite realizar estas predicciones basadas en el comportamiento de los usuarios.

Como Ayuda en la Toma de Decisiones

La ciencia de datos es empleada para proporcionar información específica que contribuye en el proceso de toma de decisiones. En estas aplicaciones, muchas veces no existe un único modelo para dar una solución final, pero los resultados obtenidos proporcionan información valiosa para los responsables de la toma de decisiones. En la investigación y análisis de crímenes, por ejemplo, se emplea minería de datos para extraer nombres de posibles sospechosos, incluyendo: registros de personas involucradas, disputas, antecedentes, elementos recuperados de la escena del crimen, ubicación satelital de los celulares de los posibles sospechosos, información de redes sociales y otros detalles. Estás estimaciones se combinan con la opinión de los investigadores para hacer seguimiento a las investigaciones y predecir posibles sospechosos. (Biron et al., 2019).

En otro ejemplo, Rajesh., et al., (2020) desarrollaron un modelo para la toma de decisiones, sobre la aprobación de préstamos, desde las perspectivas legales de los abogados y gerentes bancarios. Esta toma de decisión, se hace por medio de modelo de árbol de decisiones, en el que se ajustan pesos, a las características más importantes que se consideran al efectuar préstamos bancarios, estos pesos se fijan por medio de un sistema de predicción automatizado.

Una vez que se ha desarrollado el modelo, se puede utilizar para predecir los resultados financieros potenciales de nuevos proyectos de financiamiento, lo que permite a los inversores tomar decisiones más informadas y reducir el riesgo de pérdidas financieras. Así mismo, en la financiación de proyectos financieros Veres, O., Ilchuk, P., & Kots, O. (2021) consideran las posibilidades de utilizar métodos de ciencia de datos para decidir sobre un esquema de financiamiento de proyectos que proporcione un nivel óptimo de riesgo y rentabilidad para todas las partes interesadas.

La ciencia de datos, como ayuda en la toma de decisiones, también la podemos ver en el área del deporte, Sarlis et al., (2021) proponen una solución de ciencia de datos que ofrece información para entrenadores, científicos del deporte y la salud, gerentes y tomadores de decisiones; para reconocer las lesiones más comunes e investigar posibles patrones de lesiones durante las competiciones. Por otro lado, Claudino, J. et al., (2019) emplean la ciencia de datos para mejorar la predicción de riesgos de lesiones y el rendimiento en varios deportes de equipo.

En este proceso, se integran datos: físicos, técnicos y tácticos, junto con la experiencia experta, para informar las decisiones relacionadas con la optimización de estrategias de entrenamiento y competencia. Un aspecto destacable de estas investigaciones es el análisis y la relación entre diversos recursos de datos, que incluyen videos, análisis de tácticas, predicciones previas y mediciones de rendimiento físico mediante sistemas electrónicos en cada deportista.

Para Automatizar la Toma de Decisiones

En este caso se busca modelar con exactitud un problema específico, mediante la construcción de modelos que pueden predecir resultados futuros basados en datos históricos. Si el problema permanece constante y los modelos están entrenados y validados, pueden programarse o integrarse en sistemas informáticos para tomar decisiones automáticamente, la computadora entonces "toma la decisión". Así, la toma de decisiones se ha automatizado. Este tipo de aplicaciones se puede ver sobre todo en el control de inventarios (Farhat, J., & Owayjan, M. 2017) y en la clasificación de Correo SPAM (Alurkar, A. et al., 2017).

En aplicaciones de toma de decisiones en general, la técnica de ciencia de datos más usada es la minería de datos (Biron et al., 2019), (Schnepf, J., et al., 2022), (Rajesh et al., 2020) (Sarlis et al., 2021) (Gunawan., 2022), y sobre todo las técnicas no supervisadas debido a que no requiere una hipótesis previa sobre el valor de los datos y así se pueden descubrir patrones y relaciones. Aunque también se suelen usar técnicas de aprendizaje automático (Ebadi, A., et al., 2019) como K-Nearest Neighbor (kNN) (Li, C., et al., 2019), redes neuronales (Mompó Serrano, A., 2022) y herramientas de análisis estadístico como la optimización bayesiana secuencial, CB-SEM y regresión (Van der Voort, H., et al., 2021).

Otro aspecto para destacar, es que la herramienta más usada para el análisis de datos en estas aplicaciones suele ser Phyton (permite el uso de múltiples librerías), en menor medida se suele emplear R, SAS, KNIME, Rapid Miner, QlikView, Excel y splunk. En cuanto al modelo de análisis, suele usarse modelos de árbol de decisión (Rajeshet al., 2020; Van der Voortet al., 2021) o modelos de clasificación (Sarlis et al., 2021).

Caso de Aplicación Sector Retail

Desafíos

Aunque diferentes modelos, técnicas y procesos de extracción de patrones se han usado con anterioridad, en otros contextos de toma de decisiones, desde hace ya un largo rato, son pocos conocidos y aplicados en las micro y pequeñas empresas. En el caso específico del sector Retail: “a pesar de que este tipo de organizaciones están en la búsqueda de su fortalecimiento empresarial y de aplicar técnicas y tecnologías de avanzada, ven paradigmas que al parecer suenan difíciles y costos, bien porque no cuentan aún con una madurez en su desarrollo y modelo de negocio, porque simplemente la toma de decisiones empresariales se viene haciendo por siempre intuición de forma empírica, o la informática no es vista como un proveedor de herramientas para posibilitar y facilitar la toma de decisiones.” (Riquelme, Ruiz & Gilbert, 2006).

A medida que los minoristas diseñan sus estrategias en diversos canales, el volumen de datos internos y externos aumenta, lo que ha llevado a un crecimiento exponencial en la cantidad de fuentes de información; esto genera dificultades para comprender y extraer conocimiento de los datos, dado que: “la mayoría de los negocios de ventas al por menor enfrentan dificultades relacionadas con la recopilación, el almacenamiento, el uso, el análisis, la privacidad y la confianza en los datos.” (Al-Zahrani & Al-Hebbi, 2022).

En muchos casos, el manejo de datos de ventas, inventario, registros, tablas salariales y catálogos se lleva a cabo de manera empírica o se almacenan en un único computador de forma desordenada y sin ningún tipo de estructura, para luego acceder de forma periódica con la necesidad de extraer algún dato: “se usan por lo general sistemas contables sin plataforma propia como SG1 o similares.” (Lopez L. & Zuluaga S., 2013).

Esta mala gestión de datos conduce a información incompleta, lo que incrementa el riesgo de inconsistencias y errores, resultando en datos imprecisos que no reflejan la realidad; éstos problemas de calidad en los datos, pueden sesgar los resultados de los modelos de análisis, llevando a decisiones empresariales incorrectas. Como resultado, la gran mayoría de las veces los minoristas se encuentran en desventaja en las negociaciones con grandes distribuidores o mayoristas y cometen errores en la predicción de ventas y la gestión de inventarios, especialmente en la compra de productos perecederos, lo que genera pérdidas.

Otro desafío común en situaciones reales es la disponibilidad limitada de datos históricos, lo que puede representar un obstáculo al implementar modelos como redes neuronales ya que: "el desempeño del pronóstico depende en gran medida de tener suficientes datos históricos para el entrenamiento." (Ren, S., Chan, H. L., & Siqin, T. 2020). Finalmente, el entorno altamente cambiante en el que operan los negocios minoristas presenta un desafío significativo para la implementación de modelos de ciencia de datos. Factores internos como: los niveles de inventario, las estrategias de precios y las campañas de marketing, junto con fuerzas externas que: "no se pueden controlar, pero que pueden influir en los resultados futuros de las decisiones, como nuevas tecnologías, la presencia de nuevos competidores, cambios en la legislación o disturbios políticos" (Cabeza de Vergara & Muñoz Santiago, 2010).

Afectan de manera agresiva a los negocios minoristas, es así que la interacción entre estos factores hace que la previsión precisa sea un desafío, al ser impredecible y fluctuante, complicando la creación de modelos de ciencia de datos robustos, ya que introducen un alto grado de incertidumbre que debe ser considerado en el análisis.

Ventajas, Oportunidades y Beneficios

Según el Estudio Nacional de Emprendimiento a tenderos (Quintero, A., Medina, I., & Rodríguez-Lesmes, P., 2020), los comerciantes propietarios de tiendas de barrio formalizadas afrontan limitaciones a falta de: clientes, insumos, financiamiento, altas cargas impositivas nacionales y municipales. Como altos costos de: regulación y contratación, estos problemas ocasionan una reducción en el nivel de ingresos, menor crecimiento, dificultades en la atracción y retención de clientes y problemas en la cadena de valor, particularmente en el proceso de compras (tendero-proveedor).

Lo cual afecta el manejo del surtido, el crecimiento y la supervivencia del negocio. En este contexto, surge la oportunidad de abordar estos desafíos, mediante la identificación de las ventajas que la aplicación de la ciencia de datos puede ofrecer en la toma de decisiones en el sector Retail. Inicialmente, al analizar los datos históricos y utilizar técnicas de modelado predictivo, se pueden simular diferentes escenarios y pronosticar los resultados. De esta manera, los minoristas pueden acercarse a los resultados esperados al ajustar su estrategia y tácticas para maximizar las oportunidades de éxito, esto les permite adaptarse rápidamente a los cambios en el mercado, las preferencias del cliente y las condiciones económicas.

Por otro lado, la clasificación detallada de los clientes en diferentes grupos ofrece resultados reveladores que iluminan las preferencias y comportamientos de los clientes. Este conocimiento del cliente proporciona a los minimercados una ventaja competitiva significativa al permitir identificar en qué productos los consumidores priorizan exclusivamente el factor precio, dejando de lado consideraciones como: el prestigio de la marca, la calidad u otras características adicionales. Tambien, se pueden clasificar los productos en diferentes categorías basadas en características comunes como tipo de producto, marca, precio, ventas, estacionalidad entre otros.

(Kopap & Elfakharany, 2013), (Thomassey et al., 2003), (Jain, A., Menon, M.N., & Chandra, S., 2015). Esto les brinda una oportunidad única para adaptar su estrategia y ofrecer productos que se ajusten perfectamente a las preferencias y necesidades de sus clientes, asegurando así su satisfacción y fidelidad. Además, conocer bien a los clientes permite hacer pedidos anticipados de productos antes de que se agoten, lo que también contribuye a incrementar la lealtad del cliente.

Gracias a las técnicas y modelos utilizados en ciencia de datos, se abren diversas posibilidades estratégicas para la toma de decisiones dinámicas (cambio de un precio, variedad de productos que se tiene en venta, priorización de cuentas a pagar) y decisiones de alto riesgo como la apertura de nuevas tiendas, ampliaciones, renovaciones de tiendas y adquisiciones.

Para las decisiones dinámicas se tiene aplicabilidad en la estimación de costes (Fernández-Revuelta Pérez & Romero Blasco, 2022), el análisis del manejo del surtido, la planificación de la disposición de los productos en los estantes, análisis de fidelidad, estimación medios de pagos de los clientes, el pronóstico de ventas (Toro Ocampo et al., 2004; Ma, S., & Fildes, R., 2021) (Yelland y Dong., 2014), la gestión del inventario (Ren, S., Choi, T.M., & Liu, N., 2015) (Razmochaeva, N.V., & Klionskiy, D.M., 2019), identificar patrones de compras (Liu, H., Su, B., & Zhang, B., 2007), agrupar productos similares (Masciari et al., 2019) (Bellini, P., et al 2023) y la segmentación de clientes (Cam Gensollen, 2022; Silva Guerra, 2012).

En el futuro, se espera incluso automatizar decisiones, como realizar pedidos a proveedores, generando automáticamente órdenes de compra y eliminando la necesidad de solicitudes manuales. Esta automatización reduciría la carga de trabajo en tareas rutinarias, permitiendo dedicar más tiempo a decisiones de mayor impacto.

Por otro lado, al enfrentar decisiones de alto riesgo, como la expansión física de las empresas de comercio minorista, es crucial tomar decisiones precisas, ya estás puede influir significativamente en la sostenibilidad a largo plazo de los negocios.

En decisiones de alto riesgo, la ciencia de datos se utiliza para estimar la compatibilidad demográfica entre las empresas minoristas y sus potenciales clientes (Merino Veyl, 2015), y también permite medir el desempeño agregado de los negocios minoristas y de sus diferentes establecimientos (Gutiérrez Barrera, C. A., 2018). Emplear la ciencia de datos para la toma de decisiones, implica un cambio de paradigma en todo el negocio, porque promueve una estrategia fundamentada en datos, ya que la ciencia de datos puede ser utilizada para respaldar un análisis de FODA, al identificar riesgos y oportunidades, delimitar patrones no detectados con anterioridad, así como evaluar tendencias pasadas y futuras para llegar a una estrategia sólida y efectiva lo que impulsará a los negocios a pensar fuera de la caja, desarrollando una cultura de innovación y experimentación, tras entender el alcance de estas herramientas disruptivas.

Revisión Teórica

Toma de Decisiones

La decisión se entiende como escoger entre varias alternativas de solución, la más adecuada para alcanzar los objetivos propuestos. La relevancia y el impacto que posee la toma de decisiones en las organizaciones, han llevado a la creación de modelos que resulten eficaces para reducir la probabilidad de error, especialmente aquellos provocados por la subjetividad, con el fin de resolver los problemas de la mejor manera y en el menor tiempo posible. Los modelos de toma de decisiones: "intentan comprender, representar, describir, explicar y simular de qué forma se desarrolla el proceso de toma de decisiones organizacionales y también cómo se comportan e influyen determinados elementos intrínsecos del mismo." (Rodríguez Cruz, Y., & Pinto, M., 2018). Existen 3 tipos de modelos de toma de decisiones: descriptivos, prescriptivos y normativos.

Los descriptivos, pretenden formular una herramienta de trabajo más que una guía ideal por lo que no tiene una connotación de bueno, malo, optimo o no optimo. Los prescriptivos "se centran en lo que los decisores deben o pueden hacer para tomar decisiones y provee los mecanismos que ayudan y entrenan a las personas para tomar buenas decisiones" (Rodríguez Cruz, Y., & Pinto, M., 2018). Los modelos normativos se usan como guía, dicen cómo debe hacerse algo, proporcionan un criterio del mejor curso de acción, están orientados: "a lo que los decisores deben hacer desde una perspectiva teórica y proporcionan los procedimientos de decisión consistentemente lógicos para que a través de estos los mismos puedan decidir." (Rodríguez Cruz, Y., & Pinto, M., 2018).

En general, estos modelos de toma de decisiones proporcionan marcos y herramientas para estructurar y mejorar el proceso de decisión. "intentan establecer mecanismos, recursos y

dinámicas que faciliten los procesos de decisión indistintamente de los tipos de decisión." (Rodríguez Cruz, Y., & Pinto, M., 2018). Por otro lado, las decisiones se pueden analizar desde diferentes perspectivas, buscando entender el proceso de decisión en sí mismo. Estas perspectivas se interesan no sólo en la decisión a tomar sino la velocidad con la que se toma esa decisión, los niveles de decisión organizacional y la diferencia entre una decisión "importante" también conocida como no programada y una de "rutina" o decisión programada. (Díaz Parra, J. S. & Arango Moreno, J. F., 2020).

En el desarrollo de la actividad económica, una organización debe tomar ambas decisiones, estratégicas, que previenen sobre: acontecimientos futuros, y operativas, que dan herramientas para la gestión diaria (Holsapple et al., 2014). Las decisiones de rutina son aquellas que se toman ante circunstancias relativamente comunes, son cotidianas y repetitivas, están bien estructuradas y cuentan con suficiente información. Por otro lado, las decisiones estratégicas "casi siempre se toman en contextos de riesgo e incertidumbre" (Wilson et al., 2010, p.699) y se enfrentan a situaciones nuevas que requieren un diagnóstico y análisis para establecer causas.

Estas decisiones admiten diversas formas de solución, cada una con sus ventajas y desventajas. En estos casos, se dispone de menos información precisa, lo que puede llevar a una "sobrecarga informativa" (Bettis-Outland, 2012, p.818), desarrollándose múltiples interpretaciones sobre la información disponible, por lo que es necesario emplear la creatividad y el juicio de valor. Entre las decisiones estratégicas más destacadas que toma una empresa están "valorar las variables producto, stock, precio, tipo de cliente, o coste comercial; la planificación financiera en cuanto a las inversiones, endeudamiento y financiación; la gestión de recursos humanos, relacionado con la política de retribución o las medidas de conciliación que adoptará la empresa; y la internacionalización." (Abad-Segura, E. et al., 2022).

Para tomar una decisión, es fundamental comprender los desafíos que representan las decisiones, valorar tanto las oportunidades como las amenazas que conllevan; analizar la incertidumbre y considerar el riesgo; establecer los métodos para tomar cada una de las decisiones necesarias (Abad-Segura, E. et al., 2022). Además, como mencionan Rouhani et al. (2016) se deben considerar diversas funciones que hacen posible tomar decisiones adecuadas en un tiempo reducido. El análisis inteligente de la toma de decisiones, el razonamiento, el costo de decisión, la decisión efectiva, la ventaja competitiva y la satisfacción de la parte interesada.

Los negocios, son sistemas complejos influenciados por entornos cambiantes e inestables. Por lo tanto, al tomar decisiones, es fundamental conocer todas las características y pasos que constituyen este proceso, con el objetivo de minimizar valoraciones subjetivas. Según Machicao (2022): "es muy importante que los actores decisores tengan la información más completa posible acerca del funcionamiento del sistema y su entorno". En este sentido, la información se entiende como un recurso estratégico y se vuelve necesario para la toma de decisiones tener información: simple, fiable, oportuna, confiable, íntegra, completa, veraz, auténtica, verificable y accesible. Jansen et al. (2011, p.734) menciona que la información es un elemento fundamental en tanto: “las decisiones estratégicas tienen consecuencias importantes para el desempeño organizacional y son muchas veces el resultado de la implicación de actores desde dentro como desde fuera de la organización.”

El avance de las tecnologías y las nuevas capacidades computacionales para almacenar y procesar gran cantidad de datos con alta precisión: “permiten hacer que los datos jueguen un rol completamente distinto, ya no como herramientas de evidencia, sino como fuente de modelamiento del sustento complejo para las decisiones.” (Machicao, Jose., 2022). Según dice (Jiménez, 2017), los datos almacenados, no eran vistos como fuentes de información, para

extraer patrones de comportamiento de sus clientes, ahora esto debe ser revertidos para que genere beneficios para las organizaciones.

El análisis de estos datos puede resultar difícil, lento y consumidor de recursos cuando no puede realizarse de forma analítica (Fernández L., Romero A., 2022). Para solucionar esta necesidad, en investigaciones recientes se ha analizado el uso de la ciencia de datos en la toma de decisiones empresariales, enfocado sobre todo en la creación de herramientas de análisis apoyadas en inteligencia artificial (Fernández L., Romero A., 2022), (Chávez García., et., 2018), (Singh Yadav, N., et al., 2022), (Fernández-Revuelta Pérez, L., & Romero Blasco, Á., 2022), (Dev, M., et al., 2022).

Existen dos enfoques principales para la toma de decisiones basadas en análisis de datos. Decisiones basadas en hallazgos descubiertos a partir de datos, son: “decisiones para las cuales es necesario hacer descubrimientos dentro de los datos.” (Provost, F., & Fawcett, T., 2013) este enfoque, implica tomar decisiones informadas basadas en los hallazgos específicos obtenidos del análisis de datos. Por ejemplo, si el análisis de ventas muestra que un producto específico tiene un alto crecimiento en una región particular, la empresa puede decidir aumentar el suministro de ese producto en esa región.

Decisiones consistentes a lo largo del tiempo, son: “decisiones que se repiten, especialmente a escala masiva” (Provost, F., & Fawcett, T., 2013) Por ejemplo, un negocio puede implementar un proceso de revisión trimestral como su estrategia de marketing, basada en el análisis continuo de datos de clientes, en este enfoque: “la toma de decisiones puede beneficiarse incluso de pequeños aumentos en la precisión basados en el análisis de datos.” (Provost, F., & Fawcett, T., 2013)

Al tomar decisiones basadas en datos en el contexto empresarial, donde el margen de error es mínimo, los resultados de análisis, la aplicación de técnicas, y los modelos estadísticos y de Machine Learning, ña experiencia humana nunca debe dejarse a un lado, al tomar una decisión, sino trabajar junto con la potencia informática, "se requieren actores que tomen determinadas decisiones que no son automáticas, es decir, no dependen de un algoritmo predeterminado." (Machicao, Jose., 2022)

Tal como dicen Linoff & Berry (2011), la minería de datos permite que las computadoras hagan lo que mejor saben hacer: excavar entre una gran cantidad de datos. Eso a su vez, permite que las personas hagan lo que mejor hace, que es configurar el problema y comprender los resultados. "La reflexión, la introspección, la conciencia, la meditación, la autocrítica serán todavía durante algún tiempo muy difíciles de replicar computacionalmente y muy importantes para elevar el nivel del procesamiento de datos y de la ciencia de datos en general". (Machicao, Jose., 2022).

Por ende, es importante que los tomadores de decisión puedan confiar en los modelos estadísticos o de aprendizaje automático tratando en gran medida de tener claro cómo se obtienen los resultados de sus decisiones, "Por lo tanto, brindar información sobre los impulsores subyacentes del modelo, es imprescindible, para ayudar a personalizar las estrategias de toma de decisiones." (Coussement K., & Benoit, F., 2021).

Ciencia de Datos

La ciencia de datos puede ser definida como: "un campo interdisciplinario que utiliza esquemas científicos y algoritmos para obtener juicios e ideas a partir de los datos disponibles" (Fernández L., Romero A., 2022). "La ciencia de datos abarca a numerosos grupos de investigación de diferentes áreas (computación, estadística, matemáticas, ingeniería, etc.) que

trabajan en la propuesta de nuevos algoritmos, técnicas de computación e infraestructuras para la captura, almacenamiento y procesado de datos, etc." (García Herrero et al., 2018).

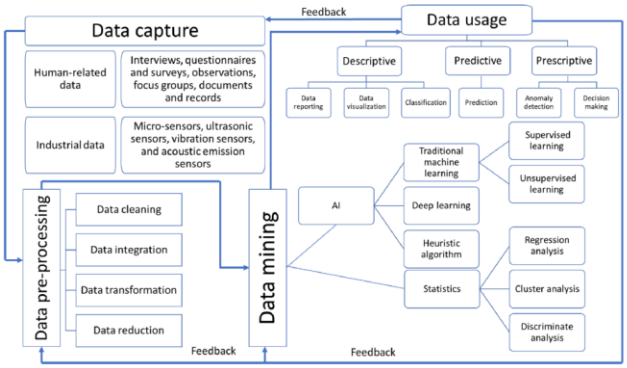
El verdadero valor de un proyecto de ciencia de datos radica en gran medida en la capacidad de identificar uno o más atributos derivados, que proporcionen información crucial sobre un problema. La interacción entre estos atributos ofrece una comprensión más profunda que si se consideraran de manera aislada. En escenarios donde la información se deriva de múltiples atributos, cada uno compuesto por dos o más variables adicionales, "la ciencia de datos nos brinda beneficios reales porque los algoritmos que utilizamos pueden, en algunos casos aprender los atributos derivados de los datos brutos." (Kelleher, J. D., & Tierney, B., 2021).

Ciclo de Vida de la Ciencia de Datos

La mayoría de los autores que abordan el ciclo de vida en proyectos de ciencia de datos (Ranjani et al., 2022; Heilman et al., 2019; Sharma et al., 2021; Karimi Dastgerdi & Javdani Gandomani., 2021; Goyal et al., 2020; Wazurkar et al., 2017) suelen dividirlo en 6 fases: descubrimiento, preparación de datos, planificación del modelo, construcción del modelo, operación y comunicación de resultados. Sin embargo, otros autores proponen diferentes números de fases según el enfoque del ciclo. Por ejemplo, (Jain, S., & Kushagra, 2022) lo describe en 5 fases, mientras que, en marcos ajustados para la extracción de información en aplicaciones empresariales, los autores utilizan 4 etapas (Singh Yadav, N., et al., 2022) (Hui, H., & Trimi, S., 2022). Estos no son ciclos lineales y las etapas no están completamente separadas, sino que requiere de muchas interacciones y a menudo volver atrás para revisar etapas anteriores.

Figura 1

Ciclo de Vida Ciencia de Datos



Nota. Fases del ciclo de vida de la Ciencia de Datos. Tomado de: Hui, H., & Trimi, S. (2022). https://doi.org/10.1016/j.techfore.2021.121242

Al comparar estos ciclos se puede ver un inicio similar, en la captura o descubrimiento de los datos. En esta fase, se incluye la adquisición de datos, la entrada de datos y la recepción de señales. La siguiente etapa es el preprocesamiento o preparación de los datos, a la que también se refieren otros autores como la limpieza, transformación de los datos, integración de los datos y normalización de los datos. Para la mayoría de los autores como Ranjani et al., (2022), el ciclo continúa con la fase de planeación del modelo, Hui, H. & Trimi, S. (2022) proponen que en esta fase se utilice la minería de datos como modelo descriptivo para un ciclo enfocado a la toma de decisiones. En esta fase también se emplean técnicas de la inteligencia de artificial o técnicas de análisis estadístico como la regresión y el clustering.

Finalmente, en todos los ciclos se tiene una capa relacionada con el uso de los datos o comunicación de resultados. En esta capa, los resultados se usan para resolver problemas como el informe de datos, la visualización, detección de anomalías y la toma de decisiones.

Modelos en la Ciencia de Datos

Un modelo es: “una estructura conceptual que interpreta los datos para generar resultados comprensibles. El modelo es una simplificación de la realidad y nunca la copia exactamente como es, un modelo siempre es imperfecto. Pero un modelo sí puede ser más próximo a la realidad que otro modelo.” (Machicao, Jose., 2022). Los algoritmos de aprendizaje automático crean modelos a partir de datos, y estos se utilizan para identificar reglas, resúmenes y patrones útiles. Estos patrones pueden representarse de diversas maneras, siendo algunas de las más populares los árboles de decisión, los modelos de regresión y las redes neuronales. Es a estás representaciones de patrones que conocemos como modelos.

La elección del algoritmo debe tener en cuenta los objetivos y las necesidades específicas del problema, así como la importancia de la interpretabilidad en relación con el rendimiento. La estructura del modelo puede ser fundamental para entender y revelar información relevante en ciertas aplicaciones, mientras que, en otros casos, los modelos se utilizan más para tareas de clasificación o etiquetado sin centrarse en los atributos específicos. Por ejemplo, en el área médica, se podría utilizar un algoritmo de aprendizaje automático, para identificar los factores que están fuertemente asociados con cierta enfermedad, proporcionando información valiosa para el diagnóstico y tratamiento. En otros casos, un modelo puede usarse como filtro de correo no deseado, cuyo objetivo principal es clasificar los correos electrónicos entrantes como spam o no spam no identificar los atributos específicos que definen un correo como no deseado.

A continuación, se revisan los modelos más empleados en la ciencia de datos y las aplicaciones para la toma de decisiones en el sector minorista.

Modelos Descriptivos

Su objetivo es resumir los datos o encontrar patrones, sin pretender predecir. Los datos se presentan como un conjunto sin estar ordenados ni etiquetados, lo que, a simple vista no ofrecen mucha información. Sin embargo, al ser procesados (limpieza, ordenamiento, transformación, visualización) estos datos pueden ser organizados de una manera comprensible, lo que permite capturar y expresar la esencia de lo que está ocurriendo. Algunos ejemplos de modelos descriptivos incluyen el agrupamiento, las reglas de asociación y el análisis correlacional.

Agrupamiento (clustering)

Es una técnica de aprendizaje no supervisado que forma grupos a partir de los datos cuando las etiquetas o clasificaciones no han sido previamente identificadas, es decir, se aplica a datos no clasificados. Al principio se desconoce la cantidad de grupos de clasificación así que se tiene 2 opciones: forzar al algoritmo o dejar que la herramienta calcule la cantidad de grupos; se suele utilizar cuando se requiere información sobre el conjunto de datos o se requiere preprocesamiento mediante algoritmos: “la principal diferencia con el algoritmo de clasificación es que es un proceso no supervisado y la clasificación es un proceso supervisado, es decir, la clasificación utiliza un conjunto de etiquetas de clase dado el conjunto de datos de entrenamiento, y se desconoce la agrupación de la etiqueta de clase” (Yu, H., Cao, L., Li, Y., & Yang, Y. 2011).

El objetivo principal del análisis de conglomerados es agrupar los puntos de datos en grupos similares bajo una misma etiqueta, “los puntos de datos idénticos se agrupan en un grupo y los diferentes se agrupan en un grupo separado” (Singh Yadav, N., et al., 2022). Esto se logra basándose en varios aspectos, como: “los métodos basados en particiones (k-medias, k-mediana y c-medias difusas), agrupamiento jerárquico (por ejemplo, aglomerativo, divisivo) y

agrupamiento basado en densidad (por ejemplo, agrupación espacial de aplicaciones con ruido basada en densidad, DBSCAN)" (Han, H., & Trimi, S., 2022).

La elección del algoritmo debe tener en cuenta el tamaño del conjunto de datos: "si son grandes o medianos, podemos usar agrupamiento basado en particiones; si los datos son pequeños, sería apropiado agruparlos jerárquicamente; Para datos ruidosos, DBSCAN sería útil" (Han, H., & Trimi, S., 2022). Una vez que los datos han sido organizados en diferentes grupos, es posible predecir la clasificación de nuevos datos que compartan características y comportamientos similares a los de los grupos existentes.

Aplicación. El clustering, se utiliza para segmentar productos en diferentes grupos basados en características similares, lo que facilita la construcción de modelos predictivos más precisos para cada grupo. Masciari et al. (2019) utilizan el algoritmo de agrupamiento en dos pasos (two-step clustering algorithm) para dividir muestras en varios subconjuntos disjuntos, haciendo que las muestras dentro de los mismos subconjuntos, sean altamente similares entre sí. Luego, los clusters, identificados se emplean para entrenar modelos XGBoost específicos para cada cluster. Al agrupar datos similares, el clustering mejora la precisión de los modelos predictivos, lo que permite reducir el exceso de inventario, minimizar los costos asociados y evitar el desabastecimiento, lo que es esencial para mantener la eficiencia operativa.

En otra aplicación, Bellini, P., et al (2023) utilizan el clustering para agrupar productos similares en función de sus características descriptivas y para segmentar a los clientes según sus comportamientos y preferencias. Según el autor, la aplicación de técnicas de clustering permite a los minoristas adaptar sus estrategias de marketing de manera más eficaz. Al clasificar a los clientes según sus comportamientos y preferencias, las campañas de marketing pueden ser más específicas y dirigidas, mejorando así su efectividad.

Correlaciones

Analiza el porcentaje de similitud entre los valores de dos variables numéricas utilizando un modelo matemático con un coeficiente de relación que toma valores entre 1 y -1. Un valor de 1 indica una relación positiva, donde al crecer una variable, la otra también crece, puede ser lineal o no lineal. Un valor de 0 indica que no hay relación. Es importante señalar que la correlación no implica causalidad; es decir, que dos variables estén relacionadas no significa que una cause el cambio en la otra. Una relación entre dos variables podría estar influenciada por una tercera variable. Por esta razón se puede combinar con modelos de regresión para estudiar relaciones entre atributos de causa-efecto.

Aplicación. El análisis de correlación puede utilizarse para reducir el número de características en un conjunto de datos en la automatización de la gestión de ventas minoristas (Razmochaeva, N.V., & Klionskiy, D.M., 2019). Cuando los datos se describen mediante un gran número de características (features), se identifican los parámetros altamente relacionados y se eliminan los redundantes, lo que ayuda a simplificar y mejorar tanto el análisis como los modelos de predicción. “Cuando se encuentran parámetros que tiene una fuerte relación lineal significa que se produce una redundancia en las características”. (Razmochaeva, N.V., & Klionskiy, D.M., 2019).

Reglas de Asociación (Association rule)

Su función principal es hallar relaciones no explicitas entre atributos categóricos, su función es la misma que las correlaciones, pero para variables nominales no numéricas. “Se utiliza en escenarios donde hay disponible un conjunto más grande de datos y se desea descubrir las relaciones o patrones, lo que permite darse cuenta de la asociación y la recurrencia entre el

conjunto de datos que reside dentro del gran repositorio de datos" (Singh Yadav, N., et al., 2022).

Las características de los datos y las asociaciones pueden cambiar con el tiempo por lo que en recientes investigaciones se han empleado reglas de asociación difusas las cuales permiten manejar datos continuos o imprecisos y expresar la relación entre elementos en términos de grados de pertenencia, por ejemplo, en lugar de decir "Si un cliente compra pan, también compra leche" una regla de asociación difusa diría "si un cliente compra una cantidad alta de pan con un grado de pertenencia de 0.8, es probable que compre una cantidad moderada de leche con un grado de pertenencia de 0.6". Se ha observado que el uso de reglas de asociación difusas, en combinación con algoritmos como Apriori, mejora la precisión en la predicción de ventas y el rendimiento en comparación con los métodos tradicionales permitiendo agrupar productos según la necesidad del cliente y mejorar la precisión en las predicciones de ventas (Ezhilarasan, C. & S, R., 2017).

Aplicación. Las reglas de asociación "infiere condiciones de valor de atributo que ocurren juntas con frecuencia en un conjunto de datos determinado" Kopap, A. H., & Elfakharany, E.-E. (2013), por lo que "podemos analizar las características de diferentes productos y las relaciones entre ellos" (Liu, H., Su, B., & Zhang, B., 2007) para identificar asociaciones como los "productos comprados con mayor frecuencia en conjunto" (Liu, H., Su, B., & Zhang, B., 2007). Esto permite ajustar los precios o crear promociones para incentivar la compra conjunta, además, tomar decisiones sobre la ubicación y estructura de los productos en las tiendas.

Por otro lado, podemos combinar "el historial de compra del cliente y la base de datos de información básica para el análisis del comportamiento de compra del cliente, encontrando los hábitos de compra de los clientes y clasificándolos." (Liu, H., Su, B., & Zhang, B., 2007) lo que

permite tomar decisiones de marketing más efectivas y dirigidas, y mejorar el enfoque en los clientes que compran ciertos productos juntos.

En general, se suele aplicar reglas de asociación en el retail para identificar patrones de compra y relaciones entre productos que puedan ayudar a mejorar las estrategias de ventas, "basándose en las reglas de asociación, las interesantes reglas y patrones de asociación ayudarán a quien toma las decisiones a tomar decisiones racionales de manera más efectiva." (Liu, H., Su, B., & Zhang, B., 2007).

Medidas de Desempeño para las Reglas de Asociación. Los algoritmos tradicionales para obtener reglas de asociación incluyen el algoritmo Apriori y el algoritmo FP-Growth. El algoritmo Apriori calcula la probabilidad de que un elemento esté presente en un conjunto de elementos dado que otro elemento este presente. El primer paso consiste en encontrar todas las combinaciones o conjuntos de elementos que cumplen con una frecuencia mínima, para luego generar reglas de asociación que involucren exclusivamente a estos conjuntos de elementos frecuentes.

Para evaluar la importancia de las reglas de asociación generadas por algoritmos como Apriori y FP-growth, se utilizan métricas como: Soporte (que tan frecuente es la regla), confianza (que tanto aparece el consecuente en las transacciones que contienen al antecedente) y la medida de elevación (Lift).

Medida de Elevación (Lift): Se considera la métrica más importante ya que es comúnmente utilizada para medir la fuerza de la relación entre dos variables y determinar si ocurren juntas con más frecuencia de lo esperado, en otras palabras, verifica que la regla no sea cuestión del azar. "La medida de elevación evalúa la cantidad de veces que A y B (es decir,

marcas, tipos y COO) ocurren juntos en comparación con la cantidad esperada de veces si fueran estadísticamente independientes" (Chiang, L. L., & Yang, C. S., 2018).

Modelos Predictivos

Los modelos predictivos buscan aproximar posibles valores futuros (predicciones basadas en probabilidad), utilizando datos en tiempo real y/o datos históricos. Los datos utilizados en estos modelos van acompañados de una salida esperada o valor objetivo, lo que permite ajustar y evaluar la precisión del modelo. Algunos modelos predictivos son:

Clasificación

"La clasificación es un enfoque de aprendizaje supervisado en el que un algoritmo aprende el mapeo óptimo de las variables de entrada en una variable objetivo (correspondiente a una clase). Después de entrenar el algoritmo, basándose en variables de entrada históricas con etiquetas de clase conocidas, el mapeo se puede aplicar a nuevas variables de entrada asociadas con datos nunca antes vistos" (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022). Cada entrada de una base de datos (instancia) pertenece a una clase (atributo clase de la instancia), el atributo busca predecir la clase desconocida de nuevas instancias. Para cada nuevo valor se espera un solo valor de salida, la salida no toma valores numéricos. "Algunos algoritmos habituales son los algoritmos de árbol, la clasificación bayesiana, la clasificación de K-vecino más cercano, etc" (Yu, H., Cao, L., Li, Y., & Yang, Y. 2011).

Clasificación suave: Se utiliza en sistemas de clasificación que requieren no solo asignar una etiqueta a cada entrada, sino también proporcionar una medida de certeza o probabilidad asociada a esa clasificación.

Categorización: No pretende el aprendizaje de una función sino el de una correspondencia. Cada elemento de un conjunto de datos puede pertenecer a múltiples categorías, lo que permite una clasificación más flexible y detallada.

Preferencias o priorización: Teniendo dos elementos de un conjunto de datos, se trata de elaborar un orden de preferencia según las características que se busquen. Por ejemplo, en medicina para decir que genes afectan más según qué enfermedades.

Parámetros de Evaluación. Para evaluar el desempeño de estos modelos, se han definido parámetros de soporte y confianza que son esenciales para el tomar decisiones.

AUC: Significa área bajo la curva. “Muestra la sensibilidad del clasificador al representar gráficamente la tasa de verdaderos positivos en relación con la tasa de falsos positivos” (Goyal, S., & Modi, N., 2017). Se puede interpretar como la probabilidad de que un modelo clasifique aleatoriamente, un positivo real por encima de un negativo real. Es decir, si se eligen al azar un positivo y un negativo, el AUC representa la probabilidad de que el modelo asigne una puntuación más alta al positivo que al negativo. “El clasificador perfecto que no comete errores alcanzaría una tasa de verdaderos positivos de 100.” (Goyal, S., & Modi, N., 2017).

Matriz de confusión: “Es una matriz de NxN, donde N es el número de clases que se van a predecir” (Goyal, S., & Modi, N., 2017). “La matriz de confusión es una forma de medir el rendimiento de la clasificación. A diferencia de otros métodos de evaluación del rendimiento, la matriz de confusión presenta un diseño de tabla de diferentes resultados de la predicción y los resultados de un problema de clasificación mientras se visualizan los resultados.” (Liço & Enesi, 2021). Tiene cuatro resultados: Verdadero Positivo, Falso Positivo, Verdadero Negativo, Falso Negativo.

Precisión: “Es la proporción del número total de predicciones que fueron correctas” (Goyal, S., & Modi, N., 2017), ya sean positivas o negativas.

Sensibilidad: “Es la proporción de casos positivos reales que fueron identificados correctamente.” (Goyal, S., & Modi, N., 2017).

Especificidad o tasa FP: “Es la proporción de casos negativos reales que fueron identificados correctamente.” (Goyal, S., & Modi, N., 2017).

Medida F: “Es simplemente una medida de combinación de precisión y recuperación.” (Goyal, S., & Modi, N., 2017).

Aplicación. Un modelo de clasificación puede ser usado en el contexto del comercio minorista, para categorizar los productos en función de sus características y demanda lo cual permite aplicar diferentes enfoques de pronóstico. Kopap & Elfakharany (2013), usaron cuatro (4) modelos de clasificación: árboles de decisión, Bayes ingenuo, red bayesiana y K-vecino más cercano (KNN) para clasificar ítems de cada marca basándose en características como: ventas totales, cantidades y área para cada ítem. Este análisis, ayudó a identificar cuáles marcas son más rentables, cuáles están descuidadas y qué áreas necesitaban más atención. También, permitió determinar si una marca estaba obteniendo resultados aceptables en áreas específicas o no.

La clasificación puede incluir datos temporales, lo que permite definir el mejor periodo de ventas para cada marca, y resulta en una mayor precisión en la predicción sobre el comportamiento futuro de las ventas. Los resultados de estas aplicaciones, muestran que es más preciso realizar pronósticos de ventas para grupos de productos, en lugar de para productos individuales. Según Thomassey et al. (2003), “se requiere un mayor número de familias de productos y criterios de clasificación pertinentes en el procedimiento de pronóstico para lograr una mayor precisión en las predicciones de ventas”.

La clasificación, también se puede emplear para elegir el mejor modelo predictivo según las características de los datos. Al clasificar los productos o las situaciones según atributos como: precio, ventas, estacionalidad, entre otros, es posible identificar el modelo que mejor se ajuste a esas características. “El resultado de la clasificación indica el mejor modelo de predicción para cada artículo. Finalmente, mediante el uso del modelo más adecuado, se logra la predicción.” (Jain, A., Menon, M.N., & Chandra, S., 2015).

Regresión

Su fin es aprender de una función, para asignar un valor de salida a una entrada de datos; evalúa la relación entre variables objetivo y variables predictoras, para prever un resultado continuo, es de prioridad reducir la diferencia entre el valor predicho y el valor real, lo que se conoce como error cuadrático. El análisis de regresión se lleva a cabo con dos objetivos: pronosticar el valor de la variable objetivo, en un escenario en el que se cuenta con información de la variable explicativa” (Singh Yadav, N., et al. 2022) el segundo objetivo es: “poder medir el impacto de la variable explicativa en la variable objetivo, es decir, identificar el factor de influencia entre las variables en función de las relaciones” (Singh Yadav, N., et al. 2022).

Los modelos de regresión suelen emplearse para: “contrastar los modelos de aprendizaje automático. La regresión es más simple y rápida que los algoritmos de aprendizaje automático, pero no pueden capturar comportamientos complejos.” (Chan, H., & Wahab, M. I. M., 2024) Algunos ejemplos de modelos de regresión son: Regresión Lineal, Regresión Polinómica y Regresión Logística.

Series Temporales

En estos métodos se utilizan los datos históricos de una variable, para generar un pronóstico del futuro, se suponen que la variable pronosticada tiene información útil para el

desarrollo del pronóstico sobre su comportamiento anterior. Uno de los desafíos principales de los modelos de series temporales es que pueden ser limitados en su capacidad para considerar y capturar todos los aspectos y variables que pueden afectar un fenómeno o proceso en el mundo real ya que, aunque “los análisis de series de tiempo son el enfoque más común para el pronóstico de la demanda en la literatura, estos métodos no pueden capturar el comportamiento no lineal de los datos” (Taha Falatouri., et al., 2022). Algunos ejemplos comunes de series temporales son: ARIMA, SARIMA, ARIMAX y Exponential Smoothing.

Aplicación. Ren, S., Choi, T.M., & Liu, N. (2015) realizaron un modelo de pronóstico de datos de panel para predecir las ventas semanales de 7 artículos de moda, encontraron que el modelo de pronóstico de datos de panel funciona mejor que el modelo tradicional de series de tiempo, ya que están involucrados más factores de impacto de la demanda. Además, mencionan que pronosticar nuevos productos o artículos es especialmente difícil en negocios minoristas, ya que a menudo carecen de datos históricos para esos productos. Por lo tanto, el uso de un modelo que tenga en cuenta múltiples factores y datos de panel puede ser más efectivo en tales circunstancias.

Técnicas Bayesianas

El teorema de bayes, permite evaluar la probabilidad de que un evento o registro pertenezca a una clase o grupo, a través de la estimación de las probabilidades condicionales inversas o a priori. Es decir, las probabilidades iniciales o previas al conocimiento de los datos, y las probabilidades condicionales inversas, que son las probabilidades actualizadas después de observar los datos. De esta manera es posible representar gráficamente la interacción entre variables e interacciones probabilísticas, por lo que suele emplearse en aplicaciones de clasificación.

Los algoritmos más empleados de este tipo de técnicas son los métodos basados en máxima verisimilitud, el algoritmo EM, redes bayesianas, el clasificador bayesiano naive y el aprendizaje profundo bayesiano el cual es: “prometedor para el aprendizaje y modelado de datos inciertos, ya que estos enfoques intentan proporcionar una interpretación razonable del resultado del aprendizaje.” (Lu, J., et al., 2019).

Aplicación. Yelland y Dong (2014) examinan la aplicabilidad de un modelo de pronóstico bayesiano para el pronóstico de la demanda de moda. Se encuentra que el enfoque bayesiano jerárquico propuesto produce resultados cuantitativos superiores en comparación con muchos otros métodos de clasificación en términos de rendimiento. En otros casos, se pueden usar estos algoritmos, para ajustar automáticamente los hiperparámetros de un modelo de aprendizaje automático, con el objetivo de mejorar su rendimiento, “se construye un modelo sustituto que se aproxima a la verdadera función objetivo en función de las observaciones acumuladas. Luego, utiliza una función de adquisición para encontrar el siguiente punto a evaluar de una región potencialmente prometedora.” (Phyu, M. M., & Khine, M. T., 2023) De esta manera, se usa el modelo sustito para explorar las combinaciones más prometedoras de manera eficiente, en lugar de evaluar directamente todas las posibles combinaciones de hiperparámetros, lo cual sería costoso.

Redes Neuronales Artificiales

Las redes neuronales artificiales, son uno de los métodos de IA más populares, ya que han demostrado ser capaces de ofrecer resultados altamente precisos en comparación con las técnicas clásicas de pronóstico en diversos dominios, incluido el comercio minorista. (Olson & Mossman 2003; Sun, Z. L., et. 2008; Zampighi et al. 2004). Las redes neuronales consisten en un conjunto de unidades básicas denominadas “neuronas” conectadas entre sí, que generalmente se

organizan en capas. Una neurona es una función en la que se hacen un conjunto muy simple de operaciones: Se toma varios valores numéricos en la entrada, se multiplican por un peso, se suman los resultados y se entrega un solo valor en la salida. Luego, la salida de la neurona pasa a través de otra función (no lineal) que se llama función de activación, culminando con la entrega de un resultado desde la capa de salida.

Para seleccionar los pesos "estas redes se entrenan mediante algún tipo de regla de aprendizaje que ajusta los pesos de las conexiones de acuerdo con los datos disponibles, tratando de minimizar una función de error apropiada" (Aburto L., & Weber R., 2007). De esta manera, durante el entrenamiento, los pesos de las conexiones entre las neuronas se ajustan repetidamente. Para hacer esto de manera efectiva, necesitan una gran cantidad de datos (ejemplos) que representen bien las variaciones y características del fenómeno que están modelando "el rendimiento de los pronósticos depende en gran medida de tener suficientes datos históricos para el entrenamiento" (Ren, S., Chan, H. L., & Siqin, T., 2020). Igualmente, según dicen Alon, I., Qi, M., & Sadowski, R. J., (2001) los métodos de redes neuronales pueden difieren en la cantidad de datos históricos necesarios, aun así, estos se necesitan para que el modelo pueda distinguir los patrones reales del ruido, evitar el sobreajuste, y generalizar bien en nuevos datos.

Con todas estas consideraciones construir una red neuronal no es una tarea fácil, implica tener una gran cantidad de datos, elegir una arquitectura apropiada (número de capas y de unidades en cada capa), seleccionar las funciones de activación, diseñar un algoritmo de entrenamiento y elegir pesos.

Otro problema de emplear este métodos es que las redes neuronales tienden a sesgarse cuando trabajan con datos de poca variabilidad. Cuando la demanda es estable, las redes

neuronales no pueden distinguir bien los patrones de demanda, “por el contrario, cuando el patrón tiene un Coeficiente de variación mayor, es más fácil que la ANN lo aprenda” (Sun, Z. L., et. 2008). Por lo que no son buenos para predecir ventas en negocios minoristas donde la mayoría de los productos tiene una demanda estable, por ejemplo, alimentos, productos de higiene personal y medicamentos.

Así mismo, si bien los métodos basados en RNA y los métodos avanzados de ANN (por ejemplo, redes neuronales evolutivas ENN) han demostrado ser capaces de proporcionar resultados con una alta precisión, “consumen mucho tiempo para realizar pronósticos debido a su utilización de algoritmos de aprendizaje basados en gradientes. Los modelos basados en ANN tomarían una cantidad sustancial de tiempo para completar una tarea básica de pronóstico de ventas (por ejemplo, puede tomar varios minutos), y las redes neuronales evolutivas (ENN) incluso tomarían más tiempo”. (Sun, Z. L., et al. 2008) “lo que se convierte en un obstáculo importante para la implementación de muchos modelos de pronóstico basados en ANN y ENN en el pronóstico de ventas en el mundo real.” (Ren, S., Chan, H. L., & Siqin, T. 2020).

Finalmente, en contexto minorista, donde las decisiones deben ser claras y justificables, las redes neuronales pueden no ser la respuesta ideal ya que son modelos complejos y a menudo son vistas como "cajas negras", esto implica que es difícil entender cómo llegan a sus conclusiones. Según Hassani et al. (2021), las redes neuronales “Son un ejemplo de técnicas de aprendizaje automático que son muy precisas, pero carecen de interpretabilidad”.

Aplicación. Los resultados de los estudios demuestran que “en promedio, las RNA obtienen resultados favorables en relación con los métodos estadísticos más tradicionales en el pronóstico de ventas minoristas” (Alon, I., Qi, M., & Sadowski, R. J. 2001) ya que permiten aproximaciones no lineales complejas entre las variables independientes y dependientes,

identificar umbrales críticos en la influencia de estas variables, y diferenciar con precisión entre movimientos ascendentes y descendentes en la variable dependiente. En este sentido al “utilizar una ANN el investigador no necesita saber el tipo de relación funcional que existe entre las variables independientes y dependientes” (Olson, D., & Mossman, C., 2003), de esta manera las redes neuronales destacan “en entornos donde las ventas están más influidas por variables exógenas como el tamaño, el precio, el color, los datos climáticos, el efecto de los medios, los cambios de precios o las campañas”. (Frank, C., Garg, A., Sztandera, L., & Raheja, A. 2003).

Aprendizaje Profundo (Deep learning)

El término "aprendizaje profundo" se utiliza para describir un subconjunto específico de redes neuronales artificiales que tienen múltiples capas (también conocidas como capas profundas) entre la entrada y la salida. Aunque técnicamente estas redes son un tipo de red neuronal, se les llama "aprendizaje profundo" debido a la profundidad de sus capas.

Es usado en contextos de grandes cantidades de datos ya que “se ejecuta de manera eficiente en diferentes escenarios cuando se trata de aprender a partir de grandes conjuntos de datos y esta es la importante ventaja del aprendizaje profundo sobre los métodos tradicionales de aprendizaje automático.” (Singh Yadav, N., et al., 2022). Algunos ejemplos de estás incluyen Redes Neuronales Convolucionales (CNNs), Redes Neuronales Recurrentes (RNNs), como LSTM o GRU, o incluso arquitecturas más avanzadas como Transformers para capturar la secuencialidad y complejidad de los datos de series temporales.

Aplicación. Ma, S., & Fildes, R., (2021) exploran un enfoque para predecir las ventas en el sector minorista utilizando un marco de meta-aprendizaje basado en redes neuronales profundas para aprender automáticamente una representación de características a partir de los datos de series temporales de ventas. El sistema meta-aprendizaje entonces utiliza estas

características aprendidas para decidir cuál de los métodos de pronóstico base (regresión, árboles de decisión, ELM) es más adecuado para cada serie temporal específica. Esto permite que el sistema sea más flexible y preciso en diferentes escenarios de ventas, ya que puede adaptar el pronóstico según el comportamiento específico de las ventas en cada caso.

Medidas de Desempeño para la Predicción de Valores Numéricos. En la previsión de la demanda, las medidas de desempeño desempeñan un papel fundamental son aplicables tanto a modelos estadísticos como a otros tipos de modelos de pronóstico. Estas medidas se utilizan comúnmente para evaluar la precisión de los pronósticos, independientemente de la técnica de modelado utilizada.

"La mayoría de los estudios evalúan el rendimiento del pronóstico mediante el error cuadrático medio (MSE), el error porcentual absoluto medio (MAPE) y el error absoluto medio (MAE)" (Ren, S., Chan, H. L., & Siqin, T. 2020). En el contexto de modelos estadísticos, estas medidas proporcionan una forma de cuantificar qué tan cerca están las predicciones del modelo de los valores reales observados en un conjunto de datos de prueba. Cuanto menor sea el valor de estas medidas, mejor será la precisión del modelo en términos de pronóstico. Además, Aburto y Weber (2007) utilizan un MSE normalizado para comparar. Ren et al. (2017) adoptan el error porcentual absoluto medio (SMAPE) para la evaluación. La medición del desempeño de todos los estudios se centra en la precisión de los pronósticos.

Sin embargo, la precisión de los pronósticos no es la única preocupación al evaluar el rendimiento de algunos métodos, se debe hacer una evaluación desde diferentes aspectos que incluyen la precisión, la velocidad, los requisitos de suficiencia de datos, la estabilidad y la facilidad de uso y otros parámetros relacionados.

Métodos y Técnicas

Un método es una forma de poner en práctica la resolución de modelos mediante el uso de herramientas específicas. Estas herramientas se emplean para implementar los métodos y resolver tareas concretas. Hay un método o técnica para cada situación y su efectividad se verá recompensada con una buena elección del mismo. Si el objetivo principal, es obtener los mejores resultados de predicción, es posible que se prefieran las técnicas de aprendizaje automático. Por otro lado, si el objetivo es comprender cómo se generaron los resultados y cómo funcionan las relaciones subyacentes, entonces los métodos estadísticos pueden ser más apropiados. "Dependiendo del problema en cuestión, es necesario determinar si están interesados en obtener los mejores resultados o en comprender cómo se produjeron esos resultados." (Hassani et al., 2021).

Técnicas Algebraicas y Estadísticas

Los métodos estadísticos se utilizan ampliamente en la predicción de ventas en negocios minoristas por tres razones principales. "En primer lugar, son fáciles de operar e implementar". En segundo lugar, corren rápido para pronosticar y calcular los resultados de la predicción (Yu, Y., et. 2011). Y finalmente, "tiene una expresión de forma cerrada que permite expresar un conjunto de datos mediante la utilización de fórmulas lo que los hace más fáciles de combinar con decisiones de operaciones comerciales (por ejemplo, gestión de inventario)" (Ren, S., Chan, H. L., & Siqin, T. 2020).

Estas técnicas son apropiadas en problemas donde: "el objetivo es comprender cómo se generaron los resultados y cómo funcionan las relaciones subyacentes" (Hassani et al., 2021), por ejemplo, para predecir productos básicos o productos con una demanda estable donde los

patrones pasados tienden a repetirse y la variabilidad es baja, debido a su capacidad para capturar tendencias y el comportamiento estacional de la demanda.

Sin embargo, en algunos casos las predicciones de demandas pueden ser afectadas por múltiples factores como las tendencias de la moda, un evento deportivo e incluso una noticia de las redes sociales, puede generar que la demanda sea irregular y volátil a lo largo de un corto periodo de tiempo. Esto lleva a que los métodos estadísticos tradicionales generalmente se vuelvan ineficaces a pesar de ser simples y rápidos (Chern et al. 2015).

En cuanto a la interpretabilidad, estos modelos pueden ser más confiables y transparentes, pero menos precisos que otros modelos. Los algoritmos de caja negra, como las máquinas de vectores de soporte o las redes neuronales (profundas) son precisos, pero no proporcionan de forma natural información sobre sus predicciones en comparación con, por ejemplo, los algoritmos basados en regresión, reglas o árboles que facilitan la interpretación del proceso de toma de decisiones.

Regresión Lineal Múltiple

La regresión lineal múltiple es un enfoque ampliamente establecido en el que se modelan relaciones lineales entre las variables explicativas (características) y la demanda. Los métodos estadísticos como la regresión lineal son altamente interpretables" (Hassani et al., 2021), lo que significa que es fácil entender cómo funcionan y por qué producen ciertos resultados. Sin embargo, los supuestos de linealidad del predictor, homocedasticidad y normalidad de la respuesta realizada en la regresión lineal a menudo se violan en el caso de patrones complejos de demanda de los clientes (Ramaekers & Janssens, 2008). Por lo que son "comparativamente bajos en términos de poder predictivo" (Hassani et al., 2021).

Aplicación. Generalmente se usa en modelos de predicción para hacer predicciones sobre valores futuros o desconocidos de la variable dependiente en función de los valores conocidos de las variables predictoras como en la predicción de ventas minoristas (Jain, A., Menon, M.N., & Chandra, S., 2015). También, se usa en modelo de forma descriptiva para comprender la relación entre parámetros de ventas y reducir la dimensión del espacio de características y mejorar la precisión y utilidad de los modelos de análisis de datos para la gestión de ventas minoristas (Razmochaeva, N. V., & Klionskiy, D. M., 2019).

Si se utiliza regresión lineal con datos no lineales, “puede resultar en problemas de baja varianza” Singh Yadav, N., et al. (2022). En estas situaciones, se sugiere el uso de regresión polinómica, ya que puede desempeñarse mejor y, por lo tanto, maximizar la complejidad del modelo. La regresión polinómica es capaz de manejar relaciones no lineales al permitir la inclusión de términos polinómicos, lo que puede mejorar la capacidad del modelo para ajustarse y capturar patrones más complejos en los datos.

Regresión Logística

Predice la probabilidad de que una observación pertenezca a una categoría o clase, permitiendo la clasificación binaria, donde los problemas se resuelven clasificando la observación en una de dos posibles categorías. En lugar de realizar una predicción continua, predice si algo pertenece o no a una categoría específica por lo que el modelo funciona bien cuando los datos se pueden separar de manera que cada categoría sea identificable de forma precisa y no se mezclen significativamente entre sí. Se utiliza cuando las salidas no tienen una relación lineal, en vez de ajustar una línea recta a los datos, se ajusta una función logística en forma de "S". Este tipo de modelo puede ser tan simple que en ocasiones “no es capaz de

capturar la complejidad de las relaciones entre diferentes factores.” (Van der Voort, H., et al., 2021).

Aplicación. La regresión logística se emplea eficazmente en problemas de clasificación binaria, donde solo hay dos resultados posibles. En el contexto del comportamiento de compra de usuarios, esta técnica puede predecir la probabilidad de que un usuario realice o no una compra (Hu, X., Yang, Y., Zhu, S., & Chen, L., 2020). Al utilizar la función sigmoide para modelar la probabilidad de ocurrencia de un evento, mapeando los resultados en un rango entre 0 y 1. Puede indicaría la probabilidad de que el usuario realice una compra con un valor cercano a 1, mientras que un valor cercano a 0 indicaría la probabilidad de que no la realice.

Análisis de Ecuaciones Estructurales (SEM)

Es una técnica de análisis estadística multivariada que permite examinar relaciones complejas entre variables observadas (medidas directamente) y variables latentes (constructos no medidos directamente, pero inferidos a partir de las variables observadas). Se podría decir entonces que SEM es, por tanto, una técnica confirmatoria más que exploratoria.

Aplicación. Al permitir comprender las relaciones entre variables latentes y observadas se puede emplear para estudiar el efecto de la heterogeneidad (diferencias individuales entre los consumidores en cuanto a los juicios que hacen y los procesos que siguen) en la formación de la satisfacción y lealtad. De esta manera ayuda a analizar y describir los diferentes segmentos de clientes basándose en sus características demográficas y hábitos de compra. Esto permite desarrollar estrategias de marketing más personalizadas y efectivas, adaptadas a las necesidades y preferencias de diferentes segmentos de clientes (Cortiñas Ugalde, M., Chocarro Eguaras, R., & Villanueva, M. L., 2010).

En una implementación (Hanaysha, J.R. 2018) utilizó el análisis SEM para analizar los factores que influyen en las decisiones de compra en la industria minorista de Malasia al relacionar los indicadores observados (variables medidas directamente) con las variables latentes o constructos teóricos. Él estudió logró demostrar que los factores examinados en el estudio (CSR, SMM, Store Environment, Sales Promotion y Perceived Value) explicaron el 72% de la varianza en la decisión de compra. Dado que el estudio muestra que el entorno de la tienda tiene un impacto positivo en la decisión de compra, los minoristas pueden invertir en mejorar la ambientación, la disposición de productos y la experiencia de compra en la tienda.

ARMA y ARIMA (AutoRegressive Integrated Moving Average)

"En el pasado se han utilizado modelos estadísticos de predicción de la demanda de la cadena de suministro como ARMA (promedio móvil autorregresivo) y modelos de regresión lineal múltiple en predicciones sistemáticas de la demanda. El problema con este tipo de modelos lineales es que sus predicciones son inexactas. Con una gran cantidad de datos y una alta potencia computacional, se puede lograr una mayor precisión incorporando una gama compleja de variables." (Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K. 2022).

El modelo ARIMA extiende el modelo ARMA al incorporar un componente adicional para manejar la no estacionariedad. Mientras que los modelos ARMA son adecuados para series temporales estacionarias (donde las características estadísticas como la media y la varianza no cambian con el tiempo), el modelo ARIMA puede tratar series con tendencias o patrones que cambian con el tiempo.

Los modelos ARIMA son especialmente útiles para series temporales estacionales, que presentan patrones regulares y repetitivos en intervalos específicos, como variaciones estacionales en la demanda de productos, "pronostica la demanda basándose en el historial de

ventas pasado y, por lo tanto, son apropiados sólo si se espera que los patrones temporales históricos continúen durante períodos de demanda futuros." (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022).

Aplicación. Los métodos de pronóstico estadístico como ARIMA se emplean a menudo por su simplicidad y facilidad operativa. La realización de pronósticos se basa en el historial de ventas pasado y, por lo tanto, son apropiados sólo si se espera que los patrones temporales históricos continúen durante períodos de demanda futuros. Por lo que pueden "lograr un rendimiento aceptable en la previsión de productos básicos con demanda estable" (Ren, S., Chan, HL. & Siqin, T. 2020).

Esto no se cumple en la mayoría de los negocios minoristas en los que "efectos exógenos, como promociones de precios irregulares, pueden tener un impacto significativo en la demanda de los clientes" (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022). Además, dada la amplia gama de unidades de diferentes productos que tienen los minoristas, "patrones de demanda muy diferentes en términos de cantidad, frecuencia, regularidad y variación, es poco probable que un solo modelo de pronóstico de la demanda pueda producir la mayor precisión de pronóstico en todos los SKU (Unidades de Mantenimiento de Stock)" (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022).

Para resolver este problema se han utilizado enfoques en los que se "trata la selección de modelos como un problema de clasificación, donde las clases corresponden a los diferentes modelos disponibles para el pronóstico" (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022), los resultados indican que no hay una solución universal que funcione para todos los casos en el pronóstico de la demanda en el sector minorista.

Otra estrategia frecuentemente adoptada es combinar este modelo con otro, por ejemplo: Aburto, Luis, and Richard Weber. (2003) combinaron modelos ARIMA y redes neuronales (NN) para pronosticar la demanda de los clientes de un supermercado chileno. En otro estudio combinaron NN y ARIMA para desarrollar un sistema inteligente híbrido para la previsión de ventas minoristas (Aburto, Luis, and Richard Weber, 2007). Wang, Y., Ye, X., & Huo, Y. (2011), utilizaron el método de K-means para clasificar 41 alimentos de un negocio minorista en 5 categorías.

Una vez que se subdividieron los alimentos en las cinco (5) subcategorías, se estructuraron modelos ARIMA separados para cada una de estas subcategorías para predecir los precios de las respectivas subcategorías. Finalmente, Masciari, E., Ji, S., Wang, X., Zhao, W., & Guo, D. (2019) proponen un modelo de pronóstico A-XGBoost que aprovecha las ventajas del ARIMA en la predicción de la tendencia de las series de datos y supera las desventajas del ARIMA aplicando XGBoost para tratar la parte no lineal de las series de datos.

En general los resultados fueron siempre mejor cuando se presentó un modelo combinado. Los enfoques en los que se combinan los modelos superan en rendimiento a los modelos puros.

ARIMAX

Es un “modelo de series de tiempo que permiten la incorporación de variables explicativas externas” (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022), como promociones de precios, en el proceso de pronóstico. Esto combina la fortaleza de los modelos de series temporales con la capacidad de considerar factores externos que afectan la demanda.

Aplicación. Arunraj, N., Ahrens, D., & Fernandes, M. (2016) demostraron que ARIMAX superó a un modelo ARIMA para predecir las ventas diarias de bananos en un negocio minorista. Esto ilustra cómo la inclusión de variables exógenas puede mejorar la precisión del pronóstico cuando existen efectos externos significativos en la demanda. La ventaja de considerar variables exógenas "permite considerar factores externos como la temperatura, la ocasión de Año Nuevo y el precio en su modelo." (Bratina, D. & Faganel,A., 2008) Aunque ARIMAX puede considerarse una herramienta valiosa para mejorar la precisión en la predicción de ventas al considerar variables externas es importante considerar la calidad del conjunto de datos ya que "el desempeño de cada método depende en gran medida de la calidad del conjunto de datos". (Dellino et al., 2015).

Técnicas de Aprendizaje Automático

Mientras el modelado estadístico es fundamental para el modelado con datos bien estructurados. "El aprendizaje automático y la IA tendrán éxito allí donde las relaciones entre los datos no se comprendan bien" (Hassani et al., 2021). Si bien son buenos para hacer predicciones, no proporcionan ecuaciones gobernantes o modelos claramente interpretables en términos del conjunto original de variables por lo cual no es fácil entender cómo llegaron a esos resultados ni qué factores influyen en las predicciones. El aprendizaje automático es una disciplina que se centra en el "entrenamiento de una máquina o sistema utilizando conjuntos de datos predefinidos para resolver problemas complejos de la última década" (Van der Voort, H., et al., 2021).

En la toma de decisiones es importante comprender los resultados de los modelos y poder interpretar el modelo y las decisiones que toman. Para esto, se pueden usar enfoques de interpretabilidad de modelos que son intrínsecos al algoritmo elegido (interpretabilidad

específica del modelo). “Recientemente varios enfoques independientes del modelo reciben mucha atención en la literatura.

Estos enfoques brindan información sobre cualquier tipo de algoritmo (de caja blanca o negra)” (Coussement K., & Benoit, F. 2021), es decir, son enfoques de interpretabilidad que son agnósticos al algoritmo utilizado (métodos de interpretabilidad independientes del modelo). Los enfoques de interpretabilidad independientes del modelo más populares incluyen: LIME, “Un enfoque que se centra en explicar las predicciones de un modelo mediante la creación de modelos locales interpretables” (Ribeiro, M. T., Singh, S., & Guestrin, C. 2016)

Valores SHAP, “Un enfoque que calcula el valor Shapley para cada característica, proporcionando una forma de medir la contribución de cada característica a una predicción” (Lundberg, S. M., Lee, S. I. 2017)

Gráficos de dependencia parcial, “Visualizaciones que muestran cómo cambia la predicción del modelo cuando se varía una característica mientras se mantienen las demás constantes” (Friedman, J.H., & Meulman, J.J., 2003).

Puntuaciones de importancia de las características de permutación, “Un enfoque que evalúa la importancia de las características al medir cómo afecta la permutación de las características al rendimiento del modelo” (Breiman, L. 2001).

Categorías Algoritmos Aprendizaje Automático. “Los algoritmos de ML se clasifican en tres categorías. Son supervisados, no supervisados y semi supervisados.” (Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S., 2018). Además, existen otros enfoques como aprendizaje por refuerzo y aprendizaje profundo. A continuación, revisaremos algunos de los algoritmos más importantes del aprendizaje automático y su aplicación en el entorno minorista:

Decisión Tree (Ábol de Decisión)

Es un algoritmo que “comprende una secuencia de decisiones para construir una estructura de árbol que cuando se aplica a un nuevo vector de características, conduce en última instancia a su clasificación” (Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R., 2022). En cada nodo, se toma una decisión a menudo binaria "sí" o "no” basada en las características de entrada, y se avanza hacia el siguiente nodo hasta llegar a un nodo hoja.

El proceso de clasificación recorre los nodos del árbol desde la parte superior o Nodo Raíz (conjunto completo de datos), tomando decisiones en cada nodo los cuales tienen dos componentes: La característica en base a la cual se toma la decisión cuya importancia se basa en su contribución a la mejora de las decisiones y el valor que define el umbral, dividiendo los datos en dos grupos (para variables binarias, esto se reduce a una decisión de sí/no).

Los nodos en el árbol, se construyen basándose en la cantidad máxima de información y estableciendo ramas del árbol según los diferentes valores. De esta manera, el modelo optimiza la clasificación mediante la maximización de la información en cada paso del árbol, asegurando que las decisiones tomadas en cada nodo sean las más informativas posibles para clasificar correctamente los ejemplos.

Este modelo tiene la ventaja de poder desarrollarse de manera interactiva esto significa que puede ser ajustado y mejorado a medida que se obtienen más datos o conocimientos. “Existen múltiples métodos para construir un árbol de decisión, con diferentes grados de interacción entre las fuentes de conocimiento. Incluso desarrollar un árbol de decisiones completamente interactivo es una opción” (Van der Voort, H., et al., 2021).

Una vez entrenados, los árboles de decisión son conocidos por ser modelos fáciles de interpretar y entender ya no solo proporcionan una clase como salida, sino que también permiten

obtener información sobre los factores que influyen en la elección del modelo (Schwartz et al., 2014). Al ofrecer explicaciones más claras sobre cómo se llegó a una decisión “los algoritmos basados en regresión, reglas o árboles facilitan la interpretación del proceso de toma de decisiones” (Coussement K., & Benoit, F. (2021).

Aplicación. Su uso en el comercio minorista suele ser para la clasificación, donde se asignan usuarios a clases particulares según su comportamiento de compra o alguna circunstancia específica. De esta manera se pueden clasificar a los clientes en función de los atributos que mejor permiten predecir el número de unidades que compran de un artículo específico (SKU). Estos atributos pueden incluir variables demográficas (como ingresos, propiedad de una casa, número de hijos, etc.) o variables relacionadas con el comportamiento de compra (como el historial de compras, frecuencia de compra, etc.)

Al utilizar la clasificación de clientes basada en árboles de decisión, se pueden realizar pronósticos de demanda más precisos (por medio de algún otro modelo como ARIMA) para cada subgrupo de clientes lo que permite mejorar las decisiones con respecto al inventario reduciendo tanto el exceso de inventario como las fallas en la venta (situaciones en las que el producto no está disponible cuando se necesita) (Bala, P.K., 2010).

Otra aplicación de este tipo de algoritmo es identificar las características (o variables) más relevantes en la predicción de ventas, descartando aquellas que no aportan valor significativo. Esto permite modelar de manera más efectiva la relación entre las variables seleccionadas y las predicciones al reducir el ruido y el exceso de complejidad (Jain, A., Menon, M.N., & Chandra, S., 2015).

En la aplicación un problema de este modelo es que puede capturar demasiado detalle de los datos de entrenamiento, incluyendo ruido o patrones específicos de ese conjunto de datos, lo

que reduce su capacidad de generalización a nuevos datos. Al ser propensos a sobre ajustarse, generalmente se restringe el crecimiento de los árboles o se podan después de su crecimiento máximo (Thomassey & Fiordaliso, 2006).

K-Means

Es un algoritmo de clustering no supervisado, que busca encontrar agrupamientos, cúmulos que se encuentran de una forma natural dentro de los datos. En el método k-means se elige de antemano el número de grupos, representados por k, “las categorías del algoritmo K-means agrupan el conjunto de elementos en un número "k" de grupos en función de su similitud. Las similitudes se calculan utilizando el método de distancia euclidiana” (Tandel, T., et al., 2020).

El objetivo es dividir los datos de entrada en k conjuntos, buscando minimizar la suma total de las distancias al cuadrado desde cada punto hasta la media del Cluster al que fue asignado. El problema aquí, es que hay muchas formas de asignar los puntos a los k clusters, por lo que encontrar la agrupación óptima es un problema muy difícil de resolver. En la mayoría de los casos en lugar de buscar la solución perfecta, se utiliza un algoritmo iterativo que generalmente encuentra una agrupación buena, aunque no necesariamente la óptima.

Aplicación. La clasificación y segmentación de datos son fundamentales para optimizar estrategias de marketing y gestión de productos en el ámbito minorista. En este contexto, los algoritmos de agrupamiento juegan un papel crucial al ayudar a las empresas a entender mejor su inventario y su base de clientes. Dando respuesta a esta necesidad el algoritmo K-means puede emplearse para clasificar y agrupar productos en diferentes categorías de rentabilidad según su tasa de beneficio. El objetivo es identificar distintos niveles de rentabilidad, lo que ayuda a categorizar los productos en grupos con rentabilidades similares, como baja, media o alta

rentabilidad (Zhang, J., & Li, J., 2016) lo que facilita una mejor toma de decisiones en cuanto a ventas e inventarios, al predecir el estado futuro de las ventas de los productos según su grupo de rentabilidad.

Por otro lado, en un estudio sobre la lealtad de los cliente en negocios minoristas (Chiang, L.-L., & Yang, C.-S. 2018), cuyo objetivo era comprender cómo los rasgos de personalidad del consumidor se relacionan con los rasgos de origen del país (COO) de las marcas de cerveza y predecir el valor del cliente (CLV). Se usó el método K-means y Fuzzy C-means (FCM), para segmentar a los clientes en diferentes categorías en función de su comportamiento de compra.

Luego se usó el enfoque RFM (Recency, Frequency, Monetary) para calcular el valor del cliente a lo largo del tiempo para cada uno de los segmentos de consumidores. Esto permitió identificar grupos de clientes con características de recencia, frecuencia y valor monetario similares. Finalmente, se examinó el índice de acierto (hit ratio) para evaluar cuántos de los clientes predichos como pertenecientes al grupo de alto valor (los mejores clientes) coincidían con los clientes reales de alto valor en el conjunto de datos.

Los resultados del modelo utilizado en el estudio tienen un buen rendimiento en la clasificación de los clientes en función de su CLV previsto, lo que puede ser útil para tomar decisiones de marketing, personalizar estrategias de marketing y tomar decisiones informadas sobre cómo interactuar con diferentes segmentos de clientes.

KNN (k-Nearest Neighbors)

Es un algoritmo de clasificación supervisada que suele utilizarse "para problemas de regresión y clasificación" (Keramati et al., 2014), y en algunos casos, también para reemplazar valores faltantes. En cuanto a la clasificación, los datos se organizan en función de su

proximidad a otros puntos cercanos. El algoritmo "Encuentra las K distancias más pequeñas entre los datos actuales y el conjunto de datos, y la salida es la etiqueta más frecuente en esos K datos" (Liço & Enesi, 2021), comparando cada nueva entrada con las instancias de entrenamiento ya conocidas. KNN no utiliza información adicional fuera de estas muestras; la clasificación "se basa únicamente en las muestras de entrenamiento, sin datos adicionales" (Kopap & Elfakharany, 2013).

El algoritmo sigue los siguientes pasos:

1. Cargar el conjunto de datos

2. Especificar un valor de K

3. Se encuentran los K puntos más cercanos al punto actual, “la clasificación k-NN encuentra un grupo de k objetos en el conjunto de entrenamiento que están más cerca del objeto de prueba y basa la asignación de una etiqueta en el predominio de una clase particular en este vecindario” (Kopap, A. H., & Elfakharany, E.-E., 2013).

4. Se devuelve la moda de las K etiquetas, Una vez que se han identificado estos k vecinos, se asigna una etiqueta al objeto de prueba en función de la clase que predomina entre estos vecinos.

Aplicación. Liço y Enesi (2021) utilizan el algoritmo de clasificación KNN para agrupar a los consumidores según sus patrones de compra y predecir sus futuras adquisiciones. Esto permite que el negocio minorista segmente mejor a sus clientes y dirija sus esfuerzos de marketing de manera más eficaz. Al comparar KNN con otros métodos de clasificación, los resultados muestran que tiene una precisión similar a la de las redes neuronales (Liço & Enesi, 2021) e incluso supera a otros modelos, como los árboles de decisión (Kopap & Elfakharany, 2013).

Aunque es un modelo muy estable y presenta una mejor precisión que otros algoritmos, KNN es considerablemente más lento cuando se trabaja con grandes volúmenes de datos (Keramati et al., 2014), pudiendo requerir hasta 60 veces más tiempo que otros algoritmos. Por esta razón, "si se busca un análisis rápido de una gran cantidad de registros, no se recomienda el uso de KNN" (Kopap & Elfakharany, 2013). Como resultado, "normalmente, el algoritmo KNN se utiliza para sistemas de recomendación, cuando el tamaño del conjunto de datos no es significativamente alto. " (Liço & Enesi, 2021).

Máquinas de Soporte Vectorial

SVM utiliza hiperplanos para separar y formar grupos o "clusters" de puntos de datos en un espacio multidimensional. “Puede haber varios hiperplanos que formen los grupos, pero se selecciona el que tenga la mayor distancia entre los puntos de datos” (Palkar, A., et., 2020). Por otro lado, “Si los datos no son separables de forma lineal, el algoritmo funciona asignando los datos a un espacio de características de mayor dimensión (donde los datos se vuelven separables)” (Ju, C., & Guo, F., 2008)

El número de hiperplanos es proporcional al número de atributos o características en el conjunto de datos. Por esto, el tiempo necesario para entrenar el modelo es una preocupación, ya que el entrenamiento con múltiples atributos puede ser lento (Palkar, A., et., 2020). Se utilizan para la clasificación y regresión, aunque cuando se aplica SVM en el contexto de regresión a menudo se denomina Support Vector Regression o SVR. Finalmente, estos algoritmos pueden ofrecer resultados de alta precisión, pero carecen de transparencia en el proceso de toma de decisiones.

Aplicación. Al incluir un mayor número de características, como la información climática, y al tener una mayor capacidad para capturar patrones complejos y relaciones no

lineales que los modelos más simples no pueden detectar, se logra capturar de manera más eficaz las relaciones complejas entre variables que podrían pasar desapercibidas con un enfoque más limitado.

Esto resulta especialmente útil para predecir las ventas diarias basándose en factores como la temperatura y la humedad. SVM (Máquinas de Soporte Vectorial) ofrece una ventaja significativa al reducir errores y mejorar la precisión de las predicciones, lo que permite a las empresas minoristas realizar previsiones más acertadas y, en consecuencia, optimizar la gestión de inventarios y las decisiones de reabastecimiento. (Chan, H., & Wahab, M. I. M., 2024).

Naive Bayes

"Es un método estadístico basado en el teorema bayesiano que se utiliza principalmente para tareas de clasificación por recuento de correlaciones de primer orden entre entradas y salidas" (Kopap, A. H., & Elfakharany, E.-E., 2013). Es útil en situaciones donde el tiempo de entrenamiento es crucial y el conjunto de datos es grande ya que puede completar el entrenamiento en "una sola pasada sobre los datos de entrenamiento, lo que lo convierte en un buen candidato para el análisis de grandes conjuntos de datos con una gran cantidad de atributos." (Kopap, A. H., & Elfakharany, E.-E., 2013).

Aplicación. Este método se utiliza exclusivamente para variables categóricas y está diseñado para tareas de clasificación. Sin embargo, "no se emplea tanto como otros métodos de clasificación" (Kopap & Elfakharany, 2013). Estos algoritmos son útiles para "extraer datos ocultos de grandes repositorios de datos sin procesar" (Kopap & Elfakharany, 2013), y han sido combinados con modelos como KNN y árboles de decisión para "clasificar productos en oferta según marca, precio y tamaño." (Kopap & Elfakharany, 2013).

A pesar de que el entrenamiento con este método tiende a ser más rápido, sus resultados suelen ser inferiores en comparación con otros modelos (Chu & Zhang, 2003; Kopap & Elfakharany, 2013). Esto se debe, en parte, a que "la brecha de precisión disminuye conforme aumenta la cantidad de datos." (Kopap & Elfakharany, 2013).

Máquina de Aprendizaje Extremo y Aprendizaje Extremo Extendido (ELME)

El algoritmo de aprendizaje extremo (ELM) es un método relativamente nuevo "para redes neuronales de alimentación directa con una sola capa oculta (SLFN)" (Sun, Z. L. et al., 2008). Las redes neuronales con una sola capa oculta son más simples que las redes neuronales profundas, que contienen múltiples capas ocultas. "En ELM, los pesos de entrada (que conectan la capa de entrada con la capa oculta) y los sesgos ocultos se seleccionan de manera aleatoria, mientras que los pesos de salida (que conectan la capa oculta con la capa de salida) se determinan analíticamente utilizando la inversa generalizada de Moore-Penrose (MP)" (Sun et al., 2008). A diferencia de los enfoques tradicionales, estos pesos no se ajustan ni se optimizan durante el entrenamiento. De este modo, ELM aprende en un tiempo mucho más corto y evita muchas de las dificultades que enfrentan los métodos tradicionales GDA y GDX, tales como los "criterios de parada, tasa de aprendizaje, épocas de aprendizaje, mínimos locales y sobre problemas sintonizados." (Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y., 2008).

Para evaluar la eficacia de estos algoritmos, se recurre al error cuadrático medio (MSE) como indicador de desempeño, además de calcular las desviaciones estándar y los coeficientes de variación para evaluar la consistencia y estabilidad de dichos algoritmos. Debido a que los pesos de entrada y los sesgos ocultos se eligen aleatoriamente, las soluciones de ELM pueden variar de una ejecución a otra. Este problema de aleatoriedad hace que sea difícil predecir exactamente cuándo una ejecución de ELM dará un buen resultado.

Para abordar esta limitación, algunos autores proponen una extensión llamada "integración de máquinas de aprendizaje extremo" (ELME). ELME es un “método de integración de regresión como una extensión del modelo ELM” (Ren, S., Chan, H. L., & Siqin, T. 2020) diseñado para mejorar la precisión en la predicción en un contexto específico, como pronóstico de ventas, En esta extensión, se ejecutan múltiples ELMs (P veces) y se calcula el promedio de las series de predicción obtenidas para reducir el error de predicción. “Aunque es más estable que ELM, todavía necesita una cantidad sustancial de tiempo para realizar predicciones, Por lo tanto, el largo tiempo de cálculo se convierte en un obstáculo importante para la implementación de muchos modelos de pronóstico basados en ANN y ENN en el pronóstico de ventas en el mundo real.” (Ren, S., Chan, H. L., & Siqin, T. 2020).

Aplicación. En el ámbito de los negocios minoristas, es posible emplear modelos como el Extreme Learning Machine (ELM) y su extensión (ELME) con el propósito de pronosticar las ventas al analizar datos históricos de ventas y factores significativos que afectan la demanda lo que facilita una toma de decisiones informada y estratégica en áreas clave como la gestión de inventarios, la planificación de producción y la implementación de estrategias de marketing.

Se ha demostrado que el rendimiento de pronóstico (efectividad) de los métodos basados en ELM para el pronóstico de ventas es mejor que muchos métodos basados en redes neuronales de retropropagación (Zhu et al. 2005; Huang et al. 2006), (Ma, S., & Fildes, R., 2021), pero “su superioridad en términos de precisión en la práctica en comparación con BPNN es, en el mejor de los casos, discutible” (Ma, S., & Fildes, R., 2021).

Además, en este tipo de algoritmos “la fluctuación en la demanda de un producto (medida mediante los coeficientes de variación) puede afectar la precisión de las predicciones” (Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y., 2008) ya que los pesos de entrada y los sesgos

ocultos (biases) se asignan de manera aleatoria al inicio del entrenamiento del modelo, lo que puede resultar en predicciones diferentes cada vez que se entrena.

En otra aplicación, Ma, S., & Fildes, R., (2021) utilizan un sistema meta-aprendizaje que utiliza la representación de características aprendidas a partir de los datos de series temporales para decidir cuál de los métodos de pronóstico base (regresión, árboles de decisión, ELM) es más adecuado para cada serie temporal específica. ELM fue uno de los modelos base que utilizaron las representaciones aprendidas para pronosticar las ventas. Los resultados mostraron que el modelo ELM "bajo la estrategia de modelado individual tienen un peor rendimiento que las regresiones ADL simples" (Ma, S., & Fildes, R. 2021), pero mejoró cuando se combinó con otros datos o métodos. Sin embargo, su efectividad sigue siendo inferior en comparación con otros modelos, como GBRT y Random Forest.

Redes Neuronales Convolucionales y Redes Neuronales Recurrentes (RNN)

Los modelos de aprendizaje profundo, como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN) más complejas, suelen tener arquitecturas más profundas y requieren grandes cantidades de datos para entrenar.

"La arquitectura general del modelo CNN siempre tiene múltiples capas convolucionales y de agrupamiento apiladas una tras otra" (Nithin, S. S. et at. 2022), lo que le permite "identificar patrones entre los pasos de tiempo" (Nithin, S. S. et at. 2022) capturando patrones temporales a corto plazo.

Aplicación. En general ha demostrado un excelente desempeño en muchos problemas de reconocimiento de patrones, y especialmente, problemas relacionados con el aprendizaje de características a partir de datos secuenciales clasificación de series de tiempo. (Zebik, Korytkowski, Angryk y Scherer, 2017; Zheng, Liu, Chen, Ge y Zhao, 2014). "Las redes

neuronales convolucionales (CNN), pueden extraer y generar automáticamente características profundas de series de tiempo, y han demostrado una fuerte robustez frente a la traducción, escalamiento y rotación de datos; esta fortaleza se deriva de tres ideas importantes que difieren de las redes neuronales tradicionales de avance; son las siguientes: campo receptivo local, intercambio y agrupación de pesos.” (Ma, S., & Fildes, R., 2021).

Finalmente, se ha aplicado en conjunto con otros modelos, como las redes de memoria a largo plazo (LSTM), "para mejorar los diseños de las redes neuronales artificiales tradicionales, incorporando capas completamente conectadas y capas de agrupación (pooling)" (Nithin et al., 2022). Lo que permite pronosticar la demanda en el sector minorista, mostrando un rendimiento superior en comparación con modelos individuales como MLP o LSTM por separado (Nithin et al., 2022).

Long-Short Term Memory (LSTM)

La red de memoria a largo plazo (LSTM) “es una categoría popular de red neuronal recurrente” (Singh Yadav, N., et al., 2022), diseñada para manejar datos secuenciales y de series temporales de manera más efectiva que las redes neuronales estándar. Se basa en RNN pero agrega conexiones de retroalimentación a diferencia de una red neuronal tradicional, compuertas de entrada, salida y olvido para seleccionar qué información recordar y olvidar. “LSTM puede identificar y recordar relaciones a largo plazo y se utiliza ampliamente para pronósticos de series temporales.” (Saha, P. et al., 2022).

Un LSTM se compone de cuatro tipos principales de capas que trabajan juntas para procesar la información a lo largo del tiempo:

- “Una capa recurrente oculta llamada bloques de memoria” (Saha, P. et al., 2022) estos bloques almacenan información durante períodos prolongados, lo que ayuda a capturar patrones en datos secuenciales.

- Unidades ocultas con puertas: Cada bloque de memoria está compuesto por unidades ocultas “Cada unidad oculta (celdas de memoria) contiene puertas para controlar el flujo de información” (Saha, P. et al., 2022), estas puertas determinan qué información debe ser almacenada, olvidada o pasada a la siguiente capa.

- Puertas de entrada y salida: “Las puertas de entrada y salida en el bloque de memoria regulan la celda que habilita el resto de la red” (Saha, P. et al., 2022), permitiendo que la red maneje y procese la información de manera efectiva.

- Puerta de olvido (forget-gate): “LSTM utiliza una puerta de olvido que ayuda a olvidar selectivamente la información de la serie temporal anterior” (Saha, P. et al., 2022), esto es importante porque en datos secuenciales, no toda la información pasada es relevante para hacer predicciones futuras.

Aplicación. Las redes LSTM se emplean en los negocios minoristas para prever la demanda de ventas. Estas redes neuronales recurrentes tienen la capacidad de identificar y recordar relaciones a largo plazo en datos secuenciales, lo que la hace especialmente efectiva para tareas que implican recordar información relevante durante períodos prolongados. Teniendo resultados con un error total menor tanto en RMSE como en MAE en comparación con ARIMA para pronósticos a largo plazo. “LSTM es más preciso que el modelo ARIMA en el escenario de predicción de más de siete días por adelantado” (Elmasdotter, A., & Nyströmer, C., 2018).

El modelo LSTM se entrena utilizando datos históricos de ventas, eventos nacionales, religiosos y deportivos en el país, y precios, permitiéndole aprender y predecir la demanda futura basada en estos datos secuenciales (Saha, P. et al., 2022).

Un modelo LSTM preentrenado en una gran cantidad de datos de series temporales puede ser ajustado para predecir eventos futuros en un conjunto de datos específico, como la demanda de productos en una tienda. Esto es posible gracias a la transferencia de aprendizaje, “una característica habitual hoy en día, ya que puede ayudar a entrenar la red neuronal profunda utilizando una pequeña cantidad de datos.” (Singh Yadav, N., et al., 2022).

Esta característica nos permite tomar un modelo preentrenado con el conjunto de datos masivo ajustarlo (o "transferir" su conocimiento) para adaptarlo a nuestro conjunto de datos más pequeño. la transferencia de aprendizaje permite a las empresas retail aprovechar modelos avanzados ya entrenados en grandes volúmenes de datos, adaptándolos a sus necesidades específicas con menos datos, lo que resulta en soluciones más precisas y eficientes para una variedad de problemas de negocio. (Singh Yadav, N., et al., 2022).

Light Gradient Boosting Machine (LGBM)

“Es un marco de aumento de gradiente basado en el algoritmo de árbol de decisión que se utiliza para clasificar” (Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K., 2022). LGBM es un algoritmo preferido para grandes conjuntos de datos debido a la naturaleza rápida del algoritmo. LGBM “utiliza una técnica de árbol de decisión, en la que las características se clasifican en contenedores y se almacenan en un histograma, después de lo cual se dividen según estos contenedores” (Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K., 2022).

Aplicación. Es aplicado en el pronóstico de ventas o predicción de la demanda en el sector minorista para prever la demanda del mercado en función de los datos de ventas históricas y mejorar la gestión de inventario en el contexto minorista, (Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K., 2022). Los resultados demuestran que “el modelo LGBM supera al modelo LSTM en términos de capacidad de pronóstico” (Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K., 2022).

Descripción Metodología

La ciencia de datos ha emergido como una disciplina esencial para la toma de decisiones en diferentes entornos, y el número de publicaciones en este campo ha ido en aumento “ya sea porque se trata de una herramienta útil en la toma de decisiones, o porque la comunidad científica lo identificó como un sistema moderno capaz de cambiar la forma en que percibimos el mundo.” (Ruiz-Lopez, F., et al., 2021).

Es fundamental que el desarrollo de estos proyectos esté guiado y estructurado, por lo cual se hace necesario el uso metodologías. Estas proporcionan un marco amplio que integra múltiples métodos y ofrece una estructura coherente para abordar problemas complejos. La mayoría de las metodologías propuestas para la aplicación de la ciencia de datos se centran en la evaluación continua, lo que permite respaldar soluciones efectivas para problemas de ciencia de datos en diversos campos.

Como señalan Ruiz-Lopez et al. (2021), “El uso de metodologías de ciencia de datos es de gran importancia, ya que son parte fundamental a partir de la cual se realizan proyectos de ciencia de datos para poder tomar decisiones basadas en datos.” CRISP-DM es una metodología ampliamente utilizada en proyectos de ciencia de datos porque se adapta a las necesidades y particularidades de cada situación. “Muestra claramente una cadena de actividades, con algunas interacciones importantes entre los pasos” (Van der Voort et al., 2021), ofreciendo un marco flexible en el que la secuencia de las fases no es estricta.

CRISP-DM abarca las siguientes etapas: comprensión de los objetivos empresariales, identificación de fuentes de datos relevantes, preparación y limpieza de los datos, análisis exploratorio, creación de modelos predictivos, y evaluación y aplicación de los resultados para la toma de decisiones informadas. Su objetivo es “facilitar el proceso de minería de datos de

extremo a extremo para el descubrimiento de conocimiento en bases de datos" (Karimi Dastgerdi & Javdani Gandomani, 2021).

Si bien existen otras metodologías usadas para guiar el proceso de análisis de datos y el desarrollo de modelos en el campo de la ciencia de datos, como KDD (Knowledge Discovery in Databases) la cual se enfoca en el descubrimiento de conocimiento en bases de datos y abarca un conjunto más amplio de técnicas y procesos (Fayyad, U. et al., 1996) y SEMMA que por su parte, está más asociada a un software en particular de SAS y se centra en la construcción de modelos y análisis de datos (Azevedo, A., & Santos, M. F., 2008)

Ninguna proporciona un marco tan completo como CRISP-DM para la toma de decisiones, la cual está principalmente orientada a la empresa y se centra específicamente en el proceso de minería de datos y descubrimiento de conocimiento a partir de los datos (Wirth, R., & Hipp, J., 2000). Teniendo como principal ventaja que está diseñada para ser independiente de cualquier software, proveedor o técnica de análisis.

Aunque CRISP-DM "propone una secuencia estructurada e iterativa de actividades, como la formulación de problemas, la consulta de datos y el modelado analítico" (Van der Voort et al., 2021), es un marco con una visión funcional "orientado hacia las actividades en lugar de hacia los actores que realizan esas actividades" (Van der Voort et al., 2021) y "rara vez consideran los intereses y las opiniones de las personas que necesitan utilizar esos resultados" (Van der Voort et al., 2021). Esto puede resultar en un modelo no tan útil o efectivo para la toma de decisiones.

Al tomar decisiones, "los trabajadores especialmente profesionales pueden tener conocimientos que potencialmente compitan con el conocimiento derivado de la inteligencia y el análisis de datos" (Van der Voort et al., 2019). Por lo tanto, no se debe depender únicamente de la analítica de datos; es esencial integrar los conocimientos derivados del análisis con la

experiencia humana, el juicio del experto o del tomador de decisiones, y su conocimiento profundo del problema y de la organización, ya que como mencionan (Molina et al., 2012), el objetivo de la toma de decisiones es encontrar soluciones óptimas en cada situación, considerando las características propias de la organización.

Es así que considerando el ciclo de vida de la ciencia de datos y apoyada en la metodología CRIPS-DM, se propone un modelo descriptivo de analítica de datos para la para la toma de decisiones en el sector Retail. Este modelo incorpora las seis etapas de la metodología CRISP-DM para guiar el proceso de análisis de datos y el desarrollo de modelos en el ámbito de la ciencia de datos, añadiendo además cinco etapas adicionales diseñadas específicamente para orientar la toma de decisiones en el contexto del negocio Retail. El modelo busca faciliten los procesos de decisión indistintamente del tipo de decisión se va a tomar (dinámica o de alto riesgo) sino que enfatiza en el proceder racional de todo el proceso, ya que "todas las decisiones siguen un proceso común, de tal manera que no hay diferencias en la toma de decisiones de tipo administrativo. El proceso de decisión puede ser descrito mediante pasos que se aplican a todas las circunstancias en las que se toman decisiones, sean estas simples o complejas" (Solano A., 2013).

A diferencia de los modelos normativos, que prescriben cómo deberían tomarse las decisiones para ser óptimas, la idea de proponer un modelo descriptivo radica en reconocer que los comercios minoristas son diversos, y las situaciones en las que se toman decisiones varían según el entorno cambiante, la gestión de datos, la cantidad de datos almacenados y el sistema utilizado para su almacenamiento. Por lo tanto, este modelo se concibe como una herramienta de trabajo, más que como una guía ideal. Su propósito es describir cómo se lleva a cabo el proceso de ciencia de datos para tomar decisiones, proporcionando un enfoque sistemático basado en

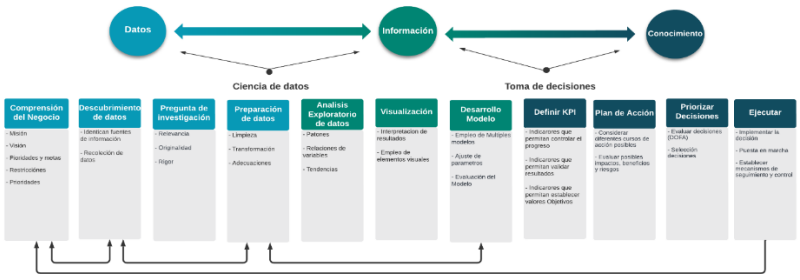
CRISP-DM, pero ajustado para considerar las opiniones y necesidades de los responsables de la toma de decisiones desde el inicio.

Desarrollo y Aplicación

En este modelo se propone iniciar en el entendimiento del negocio y elaborar una hipótesis o pregunta de negocio a la que se va a dar respuesta con el desarrollo del proyecto. Se integran etapas clave de la metodología CRISP-DM, como: la comprensión del negocio y la elaboración de hipótesis, con fases esenciales del ciclo de vida de la ciencia de datos, incluyendo la comprensión de los datos, la preparación de los datos, la creación del modelo y la evaluación de los resultados.

Figura 2

Metodología Ciencia de Datos para la Toma de Decisiones.



Nota. Descripción etapas Metodología Ciencia de datos para la toma de decisiones.

A continuación, se describen y aplican cada una de estas etapas, detallando cómo contribuyen a la implementación efectiva del modelo en un entorno real para la mejora en la toma de decisiones. Al ser una representación que se aproxima a un sistema complejo, se tienen en cuenta ciertas suposiciones y limitaciones en cada etapa.

Compresión del Negocio

Para tomar una decisión que sea acertada, es crucial hacerlo en el momento oportuno y de la manera más eficiente posible, minimizando los recursos y gastos involucrados. Esto requiere una comprensión clara del negocio, “las prioridades de la organización, sus objetivos, misión, visión, estrategia a corto, mediano y largo plazo” (Rodríguez-Cruz, Y., & Pinto, M., 2018).

Comprender el entorno implica “identificar, definir y diagnosticar lo que acontece en la organización y su contexto” (Rodríguez-Cruz, Y., & Pinto, M., (2018), “con el fin de determinar qué factores están contribuyendo al o los problemas” (Franklin Fincowsky, E. B., 2011). Una vez que se han identificado los posibles problemas, es crucial “evaluar los factores advertidos y determinar cuáles son las causas, y no tan solo los síntomas, del o los problemas reales” (Franklin Fincowsky, E. B., 2011).

Esta etapa es de suma importancia ya que, en el ciclo de vida de un proyecto de ciencia de datos, no basta con obtener resultados precisos; también “se requiere una comprensión cuidadosa del contexto cuando se trata de la evaluación de los resultados generados por la ciencia de datos, dado que es necesario darse cuenta de la forma en que se utilizarán” (Singh Yadav, N., et al, (2022) y se debe asegurar que la solución propuesta cumpla con las especificaciones y minimice el costo de oportunidad.

Además, esta evaluación es la base sobre la cual se realizan las iteraciones y ajustes en el proceso de ciencia de datos, teniendo en cuenta que “las prácticas adecuadas para utilizar son específicas del contexto y la organización” (Karimi Dastgerdi, A., & Javdani Gandomani, T., 2021), debido a los recursos disponibles, desafíos, limitaciones y necesidades particulares de cada negocio. Sin esta comprensión, las decisiones sobre qué datos utilizar, qué características destacar o qué algoritmos probar pueden ser menos efectivas, ya que, como lo señalan Karimi &

Javdani (2021), “la ciencia de datos puede ser muy iterativa e impredecible debido a sus diversas fuentes de datos, características de datos, algoritmos de modelos y arquitecturas”.

Asimismo, en el proceso de toma de decisiones, esta etapa es crucial porque implica entender el entorno en el que se generan los datos. Sin una comprensión adecuada del contexto en el que los datos fueron recolectados, es fácil malinterpretar la información, lo que a su vez puede llevar a decisiones incorrectas o menos efectivas. Como lo señalan Provost y Fawcett (2013), “el conocimiento extraído de los datos y su utilidad en la toma de decisiones dependen fundamentalmente de la aplicación en cuestión”. Van der Voort et al. (2021) destacan que “este es el paso más delicado políticamente, ya que el concepto de negocio muestra qué datos creen las partes interesadas que se deben agregar al modelo, de modo que represente el entorno organizacional correctamente”.

Finalmente, es fundamental considerar los objetivos que el negocio desea alcanzar, tales como reducir el riesgo crediticio, pronosticar los precios de las materias primas, fidelizar a sus clientes, entre otros. Ya que el objetivo empresarial es el núcleo de todo el proceso de ciencia de datos (Jain, S., & Kushagra, 2022) y “la toma de decisiones debe estar condicionada por los objetivos estratégicos de la organización y va a depender de las características del contexto en que se desarrolla” (Rodríguez-Cruz, Y., & Pinto, M., 2018).

Aplicación. En etapa se decidió utilizar una encuesta sobre diversos aspectos los negocios minoristas, incluyendo su perfil general, necesidades y desafíos, uso actual de datos, expectativas sobre la ciencia de datos y las métricas clave de éxito. La encuesta se aplicó a 8 negocios minoristas de la ciudad de Palmira Valle (Colombia) de variadas líneas de mercancía: comida, artículos deportivos, ropa y calzado. Utilizando como método de recolección la

aplicación presencial a los dueños y gerentes de los negocios usando formularios de papel. A continuación, se comparten las preguntas:

Información General del Negocio

1. nombre del negocio:

2.Tipo de comercio:

A) Ropa

B) Calzado

C) Electrónica

D) Alimentos

Otros (especificar)

3. tamaño del negocio:

A) Pequeño

B) Mediano

Identificación de Necesidades y Desafíos

4. ¿Cuáles son los principales desafíos que enfrenta su negocio actualmente? (Seleccione todas las que apliquen)

A) Encontrar y entender a los clientes adecuados para el negocio

B) Analizar qué productos o servicios están de moda

C) Comprender cómo compran y qué quieren los clientes

D) Hacer que más visitantes compren en la tienda

E) Promocionar nuevos productos o diseñadores

F) Medir el impacto de las campañas en redes sociales

G) Ofrecer productos adicionales que complementen las compras principales (ventas cruzadas)

H) Identificar a los proveedores más populares entre los clientes

Uso de Datos en la Toma de Decisiones

5. ¿Qué tipo de datos recopila actualmente su negocio? (Seleccione todas las que apliquen)

A) Historial de ventas

B) Información de clientes (edad, género, hábitos de compra)

C) Inventario de productos

D) Opiniones y comentarios de los clientes

E) Datos de marketing y campañas publicitarias

F) Datos de redes sociales

Otros

6. ¿Cómo utiliza actualmente estos datos en la toma de decisiones? (Seleccione todas las que apliquen)

A) No utiliza datos para tomar decisiones

B) Para crear estrategias de marketing

C) Para evaluar el desempeño de los proveedores

D) Para mejorar la experiencia del cliente

E) Para planificar el inventario

Otros (especificar)

Expectativas y Beneficios Esperados

7. ¿Cuáles son sus expectativas sobre el uso de ciencia de datos en su negocio? (Seleccione todas las que apliquen)

A) Incrementar las ventas

B) Mejorar la satisfacción del cliente

C) Reducir costos operativos

D) Incrementar la eficiencia del inventario

E) Optimizar campañas de marketing

Otros (especificar) _____

8. ¿Qué métricas o indicador considera más importantes para medir el éxito en su negocio? (Seleccione todas las que apliquen)

A) Volumen de ventas

B) Tasa de conversión (porcentaje de visitantes que compran)

C) Retención de clientes (clientes que vuelven a comprar)

D) Margen de beneficio

E) Tasa de retorno de productos

F) Interacción y participación en redes sociales

G) Otros (especificar)

Análisis de Respuestas. Las respuestas se analizan para obtener una comprensión integral de los negocios encuestados, lo que permite guiar el desarrollo de la solución de ciencia de datos alineada con sus necesidades y objetivos.

Nombres de negocios:

Authority FIT – Tienda de artículos deportivos

Branst Store – Venta de Ropa y accesorios

Billos – Venta de comida rápida

Dimarket – Minimercado

Viveplus – Farmacia y minimercado

La cosecha – Minimercados

Camagüey – Minimercado

Marazul – Minimercado

Análisis del Perfil de los Negocios. Tamaño y ubicación: Todos los negocios encuestados son de tamaño mediano y están ubicados en zonas residenciales, lejos de grandes centros comerciales. Su clientela está compuesta principalmente por vecinos o personas de barrios cercanos.

Número de empleados: Cada negocio tiene entre 3 y 4 trabajadores, lo que facilita una gestión directa y promueve una comunicación fluida entre los empleados.

Equipamiento y Herramientas. Caja Registradora y Lector de Código de Barras: Estos equipos son fundamentales para el registro de ventas y la gestión de inventario. Al contar con lector de código de barras, pueden almacenar información detallada de los productos vendidos, incluyendo su código, y registrar cada transacción. La caja registradora permite manejar únicamente pago en efectivo.

Características de Servicio al Cliente. Estantes para Autoservicio: Estos negocios disponen de 4 a 5 estantes en los que los productos se organizan por categorías, como bebidas, embutidos y frutas, entre otros. Sin embargo, la distribución de los productos en los estantes se realiza sin una planificación previa.

Servicio a domicilio: Los negocios ofrecen servicio de entrega a domicilio, demostrando su compromiso con mejorar la experiencia del cliente. Además, este servicio permite recopilar información (nombres, teléfonos, dirección) de los clientes que realizan compras frecuentes.

Evaluación del Uso de Datos. Historial de Ventas: La mayoría de los negocios encuestados almacena datos detallados sobre el historial de ventas. Los registros de ventas se almacenan en un único computador en tablas de Excel, lo que proporciona un sistema básico pero funcional para la gestión de ventas. Sin embargo, este método limita la capacidad de análisis debido a la falta de automatización y la dificultad de acceder a datos históricos de manera eficiente.

La información de inventario, junto con las compras a proveedores, se almacena en papel físico, lo que dificulta la consulta, almacenamiento y el acceso rápido a los datos necesarios para una toma de decisiones informada.

Planificación del Inventario: Los datos almacenados se utilizan exclusivamente para la planificación del inventario. Si bien esto es importante para asegurar que los productos estén disponibles cuando los clientes los necesiten, la falta de análisis más allá de esta función limita el potencial de crecimiento y optimización en otras áreas del negocio.

Identificación de Necesidades y Desafíos. Fidelización de Clientes: Los negocios encuestados consideran la fidelización de clientes como uno de sus principales desafíos. Esto implica un interés en desarrollar estrategias que mantengan a los clientes existentes, asegurando su regreso y aumentando el valor a largo plazo del cliente.

Aumentar la Conversión de Visitantes a Compradores: Otro desafío significativo es convertir a los visitantes en compradores. Aunque los negocios logran atraer a clientes potenciales a sus tiendas, enfrentan dificultades para persuadirlos a realizar una compra.

Falta de Uso de Redes Sociales para Publicidad: Aunque la mayoría de los negocios tiene presencia en redes sociales, no las utilizan de manera efectiva para publicidad. No consideran la promoción en redes como una necesidad primordial, lo que representa una oportunidad desaprovechada para atraer a más clientes y promover sus productos.

Ausencia de Almacenamiento y Análisis de Datos de Redes Sociales: Los negocios no almacenan datos provenientes de sus redes sociales ni miden el impacto de sus campañas.

Expectativas y Beneficios Esperados. Incrementar las Ventas: Los negocios encuestados expresan una clara expectativa de que la ciencia de datos les ayude a aumentar sus ventas. Esta expectativa refleja la necesidad de atraer más clientes y mejorar la conversión de visitantes a compradores, lo que está directamente relacionado con sus desafíos actuales.

Gestión eficiente del inventario: Otro objetivo clave es mejorar la eficiencia en la gestión del inventario para asegurar que los productos necesarios estén disponibles cuando se necesiten, minimizando costos y evitando pérdidas por exceso o falta de stock.

Descubrimiento de los Datos

Después de comprender el negocio y sus objetivos, "el siguiente paso es comprender a fondo los datos, lo que implica reunir toda la información disponible" (Jain & Kushagra, 2022). Esto es crucial, ya que "antes de poder utilizar los datos para crear un modelo, el investigador debe ser capaz de comprenderlos" (Van der Voort et al., 2021), lo que garantiza "que el investigador pueda juzgar la confiabilidad y el valor de los datos" (Van der Voort et al., 2021). Además, la ciencia de datos depende completamente de los datos para construir modelos efectivos, por lo que "una buena comprensión de los datos es esencial para el modelo que se basa en ellos" (Singh Yadav et al., 2022).

Uno de los objetivos en esta etapa es identificar las fuentes de datos que puedan proporcionar información relevante para la toma de decisiones, utilizando "el concepto de negocio para recuperar datos que las fuentes de conocimiento consideren valiosos para el modelo" (Van der Voort et al., 2021). Estas fuentes pueden incluir "la propia estrategia organizacional, la información jurídica, normativa y reglamentaria, la información económica, científico-técnica, así como las experiencias organizacionales pasadas.

A esto se suman diversas fuentes de información estratégicas internas, como indicadores de desempeño organizacional, especialistas/miembros de la organización, directivos, documentos archivísticos, bases de datos y sistemas de información. Entre las fuentes externas se encuentran sitios web e información pública de organizaciones afines, proveedores, competidores, organismos nacionales e internacionales, usuarios, consultores, decretos-ley, resoluciones y bases de datos académicas" (Rodríguez-Cruz & Pinto, 2018). Basado en este entendimiento, es posible "determinar qué datos están disponibles y cuáles se pueden utilizar para resolver el problema" (Jain & Kushagra, 2022) y también "qué pasos deben llevarse a cabo para alinear el problema comercial" (Singh Yadav et al., 2022).

Esta etapa se intenta entender y explorar los datos tal como están, "se pueden utilizar gráficos para explorar los datos; básicamente, con solo explorar los datos, puede extraer toda la información posible sobre ellos" (Jain & Kushagra, 2022), sin aún aplicar técnicas de limpieza, depuración, normalización u otros tipos de transformaciones.

Aplicación. Los datos contienen información correspondiente a ventas de familias de productos que se venden en las tienda Minorista Favorita ubicada en Ecuador, durante un periodo de 3 años, desde el 1 de enero de 2020 hasta el 31 de agosto de 2023, obteniéndose un total de 111114 observaciones y estructuran de la siguiente forma: cada fila incluye fechas,

información de la tienda y del producto, si ese artículo estaba en promoción y las cifras de ventas.

En el conjunto de datos de entrenamiento encontramos las siguientes variables:

• store_nbr identifica la tienda en la que se venden los productos.

• La familia identifica el tipo de producto vendido.

• Las ventas indican las ventas totales de una familia de productos en una tienda en particular en una fecha determinada. Los valores fraccionarios son posibles ya que los productos se pueden vender en unidades fraccionarias (1,5 kg de queso, por ejemplo, en lugar de 1 bolsa de papas fritas).

• onpromotion proporciona el número total de artículos de una familia de productos que se estaban promocionando en una tienda en una fecha determinada.

En el conjunto de datos “Festivos” se encuentra información sobre los días festivos y feriados, las fechas y el tipo de evento.

Lectura Dataset. En etapa se inicia por la lectura de los Dataset de tipo csv, la compresión de las columnas y registros almacenados.

Figura 3

DataFrame Entrenamiento

```
df_Entrenamiento.head()
```

	id	store_nbr	family	sales	onpromotion	day	month	year
0	0	1	AUTOMOTIVE	0.0	0	1	1	2020
1	1	1	BABY CARE	0.0	0	1	1	2020
2	2	1	BEAUTY	0.0	0	1	1	2020
3	3	1	BEVERAGES	0.0	0	1	1	2020
4	4	1	BOOKS	0.0	0	1	1	2020

Nota. Vista del contenido y estructura del Dataframe Entrenamiento.

En la vista del contenido, se identifican registros con ventas nulas o sin ventas, los cuales deben ser eliminados. Además, se observa la presencia de una columna id que, si bien podría ser útil en otros contextos para identificar registros únicos o realizar fusiones de datos, no es relevante para este proceso, por lo que también será eliminada.

Figura 4

DataFrame Festivos

```
df_Festivo.head()
```

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

Nota. Vista del contenido y estructura del Dataframe Festivos.

Extracción de Información sobre los Datos. A continuación, se extrae más información de los datos como el tipo de columnas que se tiene, el número de filas y la cantidad de filas nulas.

Figura 5

Información del DataFrame.

```
df_Entrenamiento.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 111144 entries, 0 to 2999501
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   id           111144 non-null  int64
 1   date         111144 non-null  datetime64[ns]
 2   store_nbr    111144 non-null  int64
 3   family       111144 non-null  object
 4   sales        111144 non-null  float64
 5   onpromotion  111144 non-null  int64
 6   month        111144 non-null  int32
 7   year         111144 non-null  int32
dtypes: datetime64[ns](1), float64(1), int32(2), int64(3), object(1)
memory usage: 6.8+ MB
```

Nota. Vista de Tipos de Datos y Número de Columnas del DataFrame.

El DataFrame entrenamiento tiene 111114 filas y 8 columnas. Las columnas “id” y “store_nbr” son de tipo int, la columna “sales” es de tipo float, y “date” y “family” son columnas de tipo categórico. No se encontraron datos nulos en ninguna columna.

Figura 6

Valores Nulos del Dataframe

```
df_Entrenamiento.isnull().sum()
```

```
id             0
date           0
store_nbr      0
family         0
sales          0
onpromotion    0
month          0
year           0
dtype: int64
```

Nota. Vista de cantidad de Valores Nulos en cada columna del Dataframe.

Figura 7

Valores Únicos del Dataframe

```
df_Entrenamiento.nunique()
```

```
id             111144
date             1684
store_nbr           2
family             33
sales           29313
onpromotion       189
month              12
year                5
dtype: int64
```

Nota. Vista de cantidad de Valores Únicos en cada columna del Dataframe.

Se extrae información sobre los nombres de las columnas y el número de valores nulos en cada una de ellas teniendo como resultado que no hay datos nulos. Además, se extrae información sobre la cantidad de valores únicos en cada columna. Ninguna de las columnas

presenta datos nulos, y todas contienen más de un valor, lo que asegura que aportan variabilidad e información relevante para el análisis. Sin embargo, se detectaron inconsistencias en la columna year, ya que se observan cinco valores distintos, cuando solo debería haber datos correspondientes a tres años.

Pregunta de Investigación o Hipótesis

Luego de tener claro en la etapa anterior, los datos con los que se cuentan "surge una de las etapas de mayor dificultad y estrechamente relacionada con las necesidades y problemáticas identificadas en el entendimiento de negocio: plantear una pregunta específica a la que se vaya a dar respuesta con el desarrollo del proyecto." (Escobar Gutiérrez, E., et al. 2021).

La ciencia de datos no se limita solo a la recolección de datos, recoger datos es sencillo, pero saber usarlos es más complicado; su esencia radica en formular preguntas científicas y estratégicas que permitan obtener insights valiosos. En este sentido, se deben plantear preguntas específicas a las que se les va a dar respuesta con el desarrollo del proyecto, como "¿qué pasó?", "¿por qué pasó?", "¿qué pasará?" y "¿qué se puede hacer con los resultados?".

Estas preguntas no solo deben alinearse con la estrategia y los objetivos establecidos por el negocio, sino también aportar nuevas perspectivas que sean factibles de investigar con los recursos de datos y técnicas a disposición.

Estás preguntas deben cumplir con tres características clave: relevancia, originalidad y viabilidad. Según Mattick, Johnston, & Croix (2018), la relevancia implica evaluar si es crucial para la entidad responder a estas preguntas; la originalidad se refiere a la capacidad de generar nueva información al responderlas; y la viabilidad asegura que las preguntas puedan ser respondidas con los datos y técnicas disponibles.

En esta etapa pueden existir dos escenarios posibles: Si se tiene una hipótesis que se quiera validar se puede establecer “una pregunta inicial que se acota a partir de aspectos geográficos, temporales, sectoriales, de contexto y de enfoque” (Escobar E., et al. 2021) para guiar el proyecto. Por otro lado, si se cuenta con datos valiosos, pero no se tiene una necesidad clara, se puede “pueden plantearse objetivos de tipo descriptivo y exploratorio para ganar un mejor entendimiento de los datos disponibles.” (Escobar E., et al. 2021)

Aplicación. ¿De cuánto serán las ventas cada mes del próximo año?

Para las predicciones de ventas mensuales para el próximo año (2024) se usarán modelos enfocados en la predicción de valores continuos con la capacidad de manejar interacciones no lineales. Cada modelo se entrenará usando datos de las ventas de los 3 años anteriores y se tendrán en cuenta las variables relacionadas con las promociones en cada año, los días festivos, y la familia de productos. Se espera que estas variables proporcionen información clave sobre la dinámica de ventas y permiten a los modelos aprender los patrones históricos que se utilizan para estimar las ventas del próximo año.

¿Cuántas serán las ventas cada mes si se aumenta en un 10% las promociones?

Para esta predicción se considera la suposición de un incremento del 10% en las promociones de cada año. Las nuevas predicciones, que reflejan el incremento en las promociones, se compararán con las ventas reales de los años anteriores y con las predicciones originales sin el aumento de promociones a través de un gráfico, lo que facilitará el análisis del impacto a lo largo de los meses en 2024.

¿Cuántas serán las ventas mensuales de cada familia de productos? y ¿Qué mes se vende más cada producto?

Para identificar el mes con más ventas para cada producto, se agruparán los datos por mes y familia de productos, y luego se observarán las ventas predichas. Para esto se usará un gráfico interactivo que se genera en la aplicación Dash, el cual permitirá visualizar las ventas mensuales predichas por cada familia de productos. Así, se podrá observar las ventas tanto bajo condiciones de promociones normales como con un incremento del 10% en las promociones. De esta manera, es posible identificar no solo las ventas estimadas para cada mes, sino también determinar en qué mes se esperan las mayores ventas para cada familia de productos.

¿Qué día de la semana de cada mes se vende más cada familia de productos?

Para responder a esta pregunta, se filtrarán los datos por familia y mes, para luego agrupar las ventas por día de la semana. Para la visualización se utilizará un gráfico interactivo que permita visualizar las ventas predichas para cada día de la semana, lo que facilitará la identificación de los días de la semana con más ventas para cada familia de productos.

Preparación de Datos

En esta fase, ya se "han examinado detenidamente los datos, de modo que sabemos qué datos están disponibles y qué significan" (Van der Voort et al., 2021). Ahora, "estos datos deben adecuarse o procesarse de forma que sean aptos para las etapas de exploración, modelamiento y análisis." (Escobar Gutiérrez, E., et al. 2021). Si el análisis no se inicia con datos confiables, nos daremos cuenta de que los resultados carecen de importancia o validez.

Estás adecuaciones y transformaciones pueden ser: "identificar los datos relevantes, fusionar los conjuntos de datos para integrarlos, limpiarlos, tratar los valores faltantes eliminándolos o imputándolos, eliminar los datos erróneos y comprobar los valores atípicos con diagramas de caja y manejarlos son todas partes de este proceso.

Crear datos nuevos y extraer nuevas características de los datos existentes. Eliminar las columnas y características superfluas de los datos y formatearlos de acuerdo con las especificaciones." (Jain, S., & Kushagra., 2022. "También se analizan los problemas de calidad de los datos. Por ejemplo, es posible que falten valores, en este caso se debe idear una estrategia para abordar este tipo de problemas" (Van der Voort et al., 2021).

Estas transformaciones se realizan "con el objetivo de contar con información que se pueda considerar pertinente, necesaria y suficiente para tomar las decisiones, con criterios satisfactorios asociados a la calidad, cantidad y forma, permitiendo que tenga las siguientes cualidades: veraz, íntegra, auténtica, confiable, simple, completa, verificable, oportuna y accesible, entre otros". (Rodríguez-Cruz, Y., & Pinto, M., 2018).

"Otro aspecto importante es que, durante esta actividad, los datos de validación y prueba se reservan, de modo que el modelo pueda evaluarse en actividades posteriores" (Van der Voort et al., 2021). Esto resalta la importancia de esta etapa, siendo "la preparación de los datos el paso más importante y que más tiempo requiere en todo el ciclo de vida". (Jain, S., & Kushagra., 2022), debido principalmente a que "la calidad del modelo estará determinada por los datos que se proporcionen" (Jain, S., & Kushagra., 2022) y "la IA aún no está lo suficientemente avanzada como para identificar anomalías en los datos." (Hassani et al., 2021).

Aplicación. Revisión del tipo de datos: El código comienza asegurándose de que la columna date esté correctamente formateada como datetime. Este paso es esencial para que las operaciones basadas en el tiempo (como el agrupamiento o la extracción de componentes de fecha) puedan llevarse a cabo sin problemas. La comprensión correcta del formato de las fechas es vital para cualquier análisis temporal, como la predicción de ventas basada en datos históricos.

Eliminación de datos faltantes y duplicados: Se eliminan las filas en df_Entrenamiento donde la columna sales tiene un valor de 0 (ya que se consideran ventas no válidas).

Después, se eliminan las columnas que contienen valores nulos (NaN) con dropna(axis=1).

Se eliminan las filas duplicadas del DataFrame df_Entrenamiento utilizando drop_duplicates(inplace=True).

Creación de nuevas características: Se crean nuevas columnas a partir de la columna date en df_Festivo y df_Entrenamiento, extrayendo el día (day) y el nombre del mes (month_name). Esto permite analizar las ventas y promociones por día o mes.

Además, se crea la columna Holiday en df_Festivo, marcando los días festivos (True o False), lo que será útil para identificar los impactos de los festivos en las ventas.

Identificación y eliminación de columnas irrelevantes: Se revisan las columnas categóricas como store_nbr, family, y onpromotion para asegurarse de que contengan más de una categoría. Esto garantiza que las columnas sean útiles para el análisis.

Se eliminan columnas irrelevantes o no necesarias (por ejemplo, transferred, locale_name, description, etc.) en df_Festivo para reducir el ruido en el conjunto de datos.

Corrección de errores tipográficos en variables categóricas: Se crea un diccionario de reemplazos para corregir los nombres de las categorías en la columna family del DataFrame df_Entrenamiento. Esto estandariza los valores y facilita un análisis más coherente.

Agregación de datos por año: Se reorganiza el dataframe de forma que se agrupan los datos de ventas y promociones por día, mes y familia para cada año (2020-2023). Luego, se renombran las columnas de acuerdo con el año específico (sales_2020, onpromotion_2020, etc.).

Los datos de cada año se combinan en un único DataFrame (df_combined), lo que facilita el análisis comparativo de ventas y promociones entre diferentes años.

Integración de festivos: El DataFrame resultante df_combined se fusiona con los datos de festivos (df_Festivo) para añadir una columna que indique si un día es festivo o no (Holiday). Se rellena con False para los días no festivos.

Creación de ventanas móviles (Rolling Windows): Se añade una columna en el que se incluyen los promedios móviles de las ventas para los últimos 7 días (ventanas móviles) para cada año. Esto ayuda a capturar tendencias a corto plazo y estacionalidades en los datos de ventas, lo cual es útil para el análisis y la predicción.

Figura 8

Dataframe Final

	day	month_name	family	sales_2020	onpromotion_2020	sales_2021	onpromotion_2021	sales_2022	onpromotion_2022	sales_2023	onpromotion_2023	Holiday	sales_2020_rolling_7	sales_2021_rolling_7	sales_2022_rolling_7	sales_2023_rolling_7
0	1	February	Accesorios Automoviles	16.000	0.0	14.375000	0.0	13.5000	0.0	10.000	0.0	False	16.000	14.375000	13.5000	10.000
1	1	February	Alimentos Congelados	411.314	0.0	543.326125	0.0	372.9735	4.0	254.263	4.0	False	411.314	543.326125	372.9735	254.263
2	1	February	Alimentos preparados	409.900	0.0	283.263687	0.0	458.1510	0.0	231.573	0.0	False	409.900	283.263687	458.1510	231.573
3	1	February	Ave de corral	1130.998	0.0	1430.658937	0.0	1661.1150	1.0	946.626	0.0	False	1130.998	1430.658937	1661.1150	946.626
4	1	February	Bebidas	4500.000	0.0	4584.187500	0.0	4407.0000	3.0	5648.000	51.0	False	4500.000	4584.187500	4407.0000	5648.000
5	1	February	Belleza	4.000	0.0	10.062500	0.0	16.5000	1.0	7.000	0.0	False	4.000	10.062500	16.5000	7.000
6	1	February	Carnes	2404.742	0.0	1332.653063	0.0	967.8045	0.0	906.343	0.0	False	2404.742	1332.653063	967.8045	906.343
7	1	February	Comestibles l	9384.000	0.0	9618.312500	0.0	10465.5000	9.0	7194.000	82.0	False	9384.000	9618.312500	10465.5000	7194.000
8	1	February	Cuidado Personal	608.000	0.0	688.562500	0.0	981.0000	0.0	486.000	0.0	False	608.000	688.562500	981.0000	486.000
9	1	February	Decoración	58.000	0.0	115.000000	0.0	96.0000	0.0	39.000	0.0	False	58.000	115.000000	96.0000	39.000

Nota. Vista del Dataframe luego de la preparación y limpieza de datos.

Análisis Exploratorio de Datos

Es importante que la fase anterior se realice con detenimiento ya que para el análisis es necesario que los datos estés preparados y limpios, “la calidad, la cantidad y la granularidad de los datos disponibles determinará, en gran medida, cómo estos pueden ser aprovechados y qué tipo de técnicas y modelos pueden aplicarse. De acuerdo con el tipo de datos —estructurados,

semiestructurados, no estructurados, secuenciales, geográficos, etc.— también habrá unas técnicas de procesamiento, análisis y modelamiento más apropiadas que otras" (Escobar Gutiérrez, E., et al. 2021).

Este paso se realiza antes de desarrollar el modelo real, ya el EDA permite "adquirir una noción general de la solución y los elementos que la influyen." (Jain, S., & Kushagra., 2022) al entender el contenido de los datos, comprender cuáles son las variables más importantes y cómo estas se relacionan entre sí, "analizar en profundidad cada aspecto de las características para obtener algunos patrones que nos ayuden a modelar" (Abbas, W., Usman, M., & Qamar, U., 2022) e identificar datos atípicos o valores anómalos, que pueden requerir un análisis más detallado o la eliminación de esos puntos del conjunto de datos.

Es necesario abordar la incertidumbre y estar preparado para descubrimientos inesperados durante el proceso de análisis de datos ya que "el análisis exploratorio de los datos disponibles en esta etapa puede llevar a descubrimientos contraintuitivos." (Alexander, D.T., & Lyytinen, K.J., 2017).

Aplicación. Descripción estadística: Se usa la función, describe () para obtener un resumen estadístico de las variables numéricas, incluyendo la media, desviación estándar y rangos, lo que ayuda a entender la distribución general de los datos.

Figura 9

Descripción Estadística Dataframe

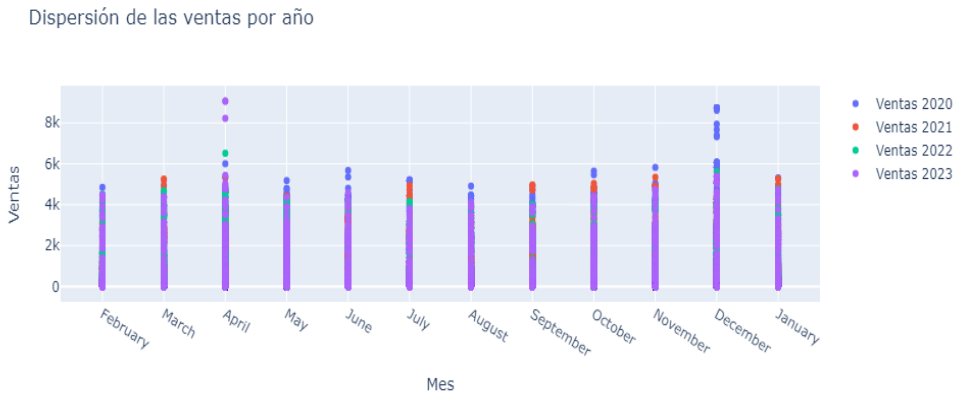
```
df_combined.describe()
```

	day	sales_2020	sales_2021	onpromotion_2021	sales_2022	onpromotion_2022	sales_2023
count	18464.000000	12388.000000	15445.000000	15445.000000	16396.000000	16396.000000	18008.000000
mean	15.326365	1289.032712	1018.487294	2.887990	932.390088	4.753111	894.008898
std	8.903751	2333.851948	1914.087782	21.413536	1914.627978	25.194188	2154.460126
min	1.000000	2.000000	1.250000	0.000000	1.000000	0.000000	1.000000
25%	7.000000	44.000000	21.250000	0.000000	18.000000	0.000000	16.000000
50%	15.000000	455.880000	248.078750	0.000000	185.000000	0.000000	90.555500
75%	24.000000	1346.024000	985.666250	0.000000	827.598000	2.000000	668.640245
max	31.000000	21308.000000	11016.250000	377.000000	35917.000000	438.000000	127587.000000

Nota. Descripción Estadística de las columnas numéricas del Dataframe.

Figura 10

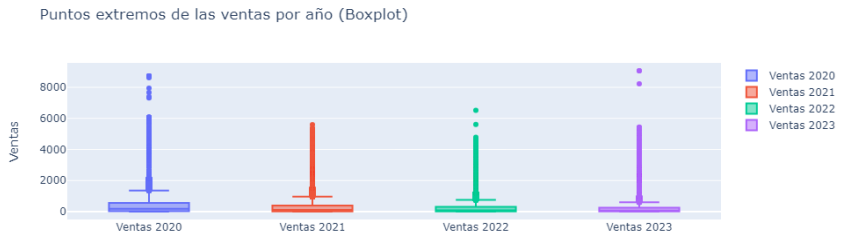
Análisis de Dispersión de las Ventas por Mes



Nota. Dispersión de las ventas por mes a lo largo de varios años.

Figura 11

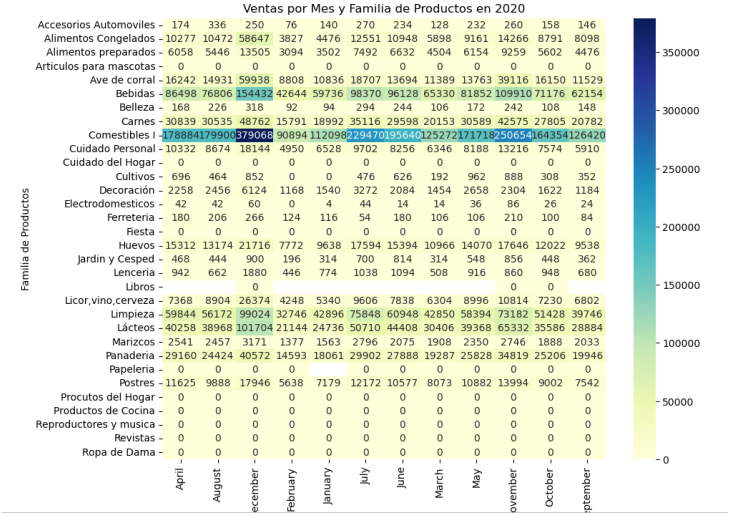
Puntos Extremos de las Ventas por Año



Nota. Visualización de puntos extremos y atípicos de las ventas de cada año.

Figura 12

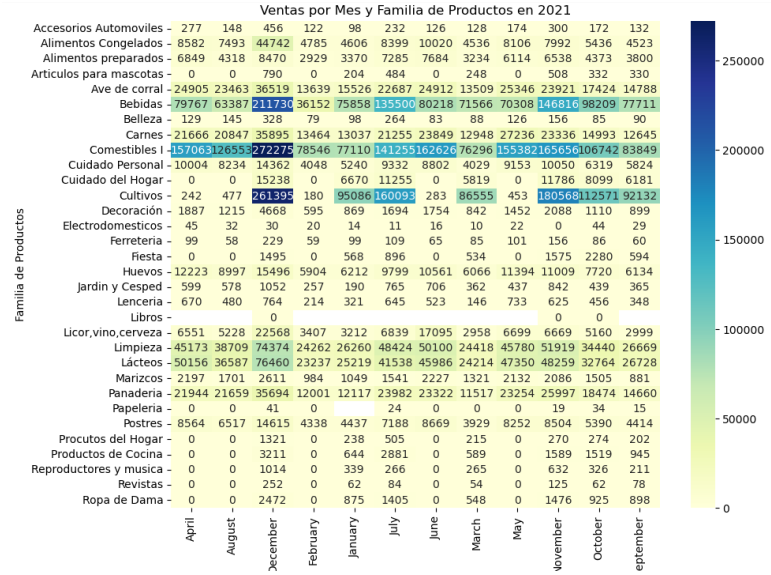
Mapa de Calor Año 2020



Nota. Mapa de calor entre las familias de productos y las ventas en cada mes del Año 2020.

Figura 13

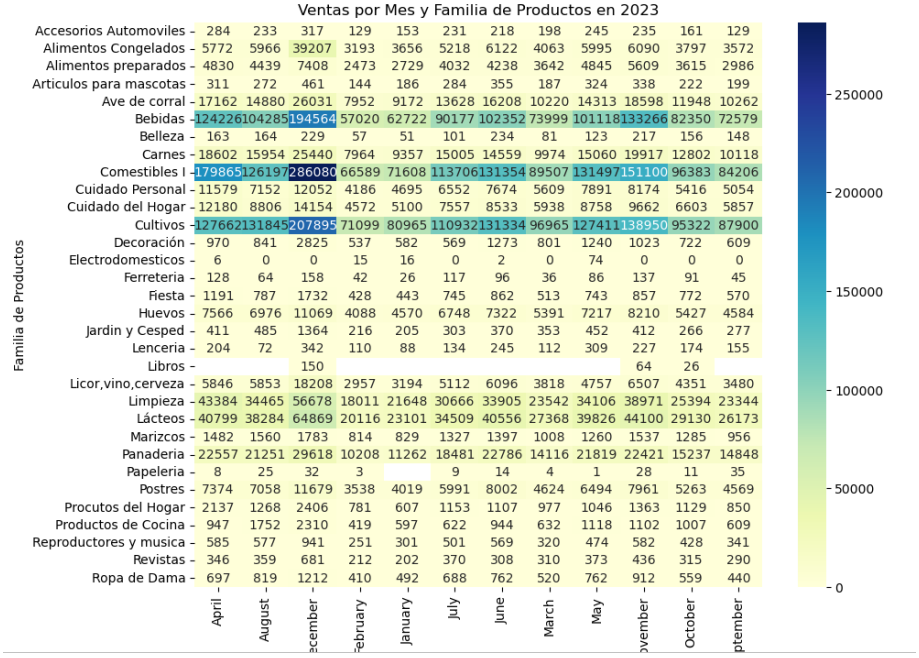
Mapa de Calor Año 2021



Nota. Mapa de calor entre las familias de productos y las ventas en cada mes del Año 2021.

Figura 14

Mapa de Calor Año 2023



Nota. Mapa de calor entre las familias de productos y las ventas en cada mes del Año 2022.

Análisis de dispersión: Se visualiza la dispersión de las ventas por mes a lo largo de varios años. Esto ayuda a identificar tendencias estacionales o fluctuaciones anuales en las ventas.

Visualización de extremos y puntos atípicos: Se utilizan diagramas de caja para resaltar los puntos atípicos en las ventas de cada año. Los valores atípicos pueden reflejar días con ventas inusuales, lo que podría requerir un ajuste.

Análisis de correlación: Se usa un mapa de calor para analizar la correlación entre meses y familias de productos para cada año. La visualización muestra la magnitud de las ventas a lo largo de las familias y los meses, lo que permite detectar patrones y tendencias de temporada.

Visualización

"Con la ayuda de la visualización de datos, podemos acceder a grandes cantidades de datos y visualizarlos" (Ranjani et al., 2022) lo que permite "investigar cada componente individualmente e integrarlo con otras características" (Jain & Kushagra, 2022) facilitando así la detección de tendencias y anomalías. Además, nos permite "explorar datos de diversas fuentes para determinar qué datos son relevantes para fines predictivos" (Dhar, V., 2013) y mostrar los "datos en un contexto visual para que se pueda comprender fácilmente su significado" (Singh & Saxena, 2021), lo que permite generar informes no solo visualmente atractivos, sino que también proporciona información contextual que respalda la toma de decisiones.

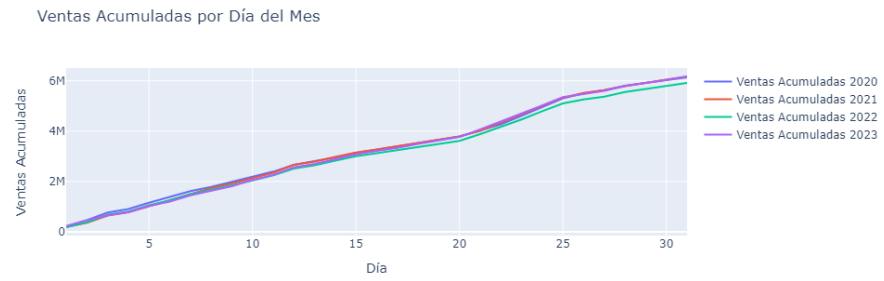
Es fundamental garantizar que el tipo de visualización utilizado ilustre de manera óptima el patrón, problema o tendencia que se desea destacar. Por ejemplo, "los gráficos de barras se

utilizan para visualizar la distribución de los datos dentro de las diferentes variables de características, y los diagramas de dispersión y los mapas de calor se utilizan para visualizar las relaciones entre las diferentes características". (Jain, S., & Kushagra., 2022).

Aplicación. Ventas acumuladas: Se observa la evolución diaria de las ventas acumuladas a lo largo de cada año, lo cual permite identificar periodos clave de aumento, como durante promociones o eventos especiales. En este caso, se identifica una tendencia de crecimiento casi lineal y constante, lo que sugiere que no existen picos significativos en días específicos, ni un incremento notable en ningún periodo particular.

Figura 15

Ventas Acumuladas por Día del Mes



Nota. Visualización ventas acumuladas por día del mes para cada Año

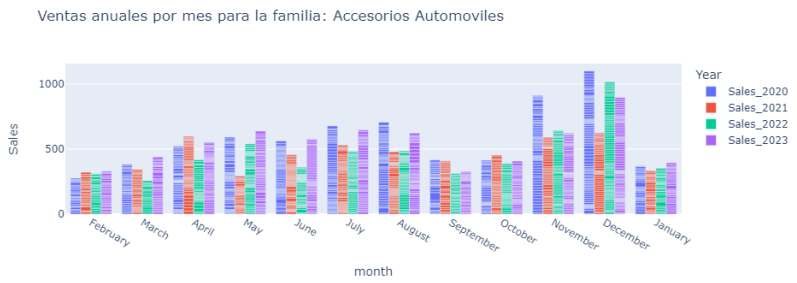
Figura 16

Ventas Mensuales de Familia de Productos "Limpieza"

Nota. Gráfico de barras que muestra las ventas mensuales de la familia de productos "Limpieza"

Figura 17

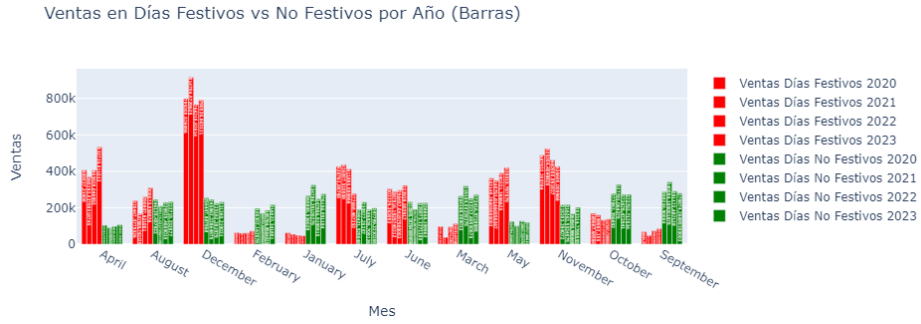
Ventas Mensuales de Familia de Productos "Accesorios Automóviles"



Nota. Gráfico de barras que muestra las ventas mensuales de la familia de productos "Limpieza"

Figura 18

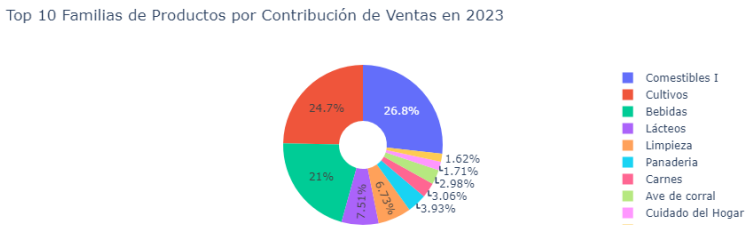
Ventas Días Festivos vs Días No Festivos



Nota. Comparación de ventas en días festivos y no festivos para cada Año

Figura 19

Contribución de Ventas de Cada Familia de Productos



Nota. Porcentaje de Contribución de Ventas de Cada Familia de Productos Año 2023

Modelado

Para este punto, en el que se han analizado los datos y se han encontrado patrones, los datos se han convertido en Información: Conjunto de datos organizados que al leerlos nos produce un significado. De esta manera se pueden dar solución a preguntas de investigación más específicas. Una vez se entienden los datos, se construyen y evalúan modelos para resolver el problema de análisis de datos específico que se está abordando. No siempre son modelos de predicción, se pueden usar por ejemplo modelos de clasificación de datos. “El modelado de datos es el núcleo del análisis de datos. Dependiendo de si el problema es de clasificación, regresión o agrupamiento, esta fase comprende la selección del tipo de modelo apropiado.

Debemos seleccionar e implementar cuidadosamente los algoritmos a implementar después de seleccionar la familia de modelos y los diversos algoritmos dentro de esa familia, como también, afinar los hiperparámetros de cada modelo para lograr el rendimiento deseado. También es importante lograr el equilibrio adecuado entre rendimiento y generalización. No queremos que el modelo aprenda los datos y falle cuando se enfrente a nuevos datos” (Jain, S., & Kushagra. 2022).

Selección Modelo

A la hora de implementar modelo de ciencia hay ciertos problemas que se deben considerar “ya que cada algoritmo tiene sus fortalezas y debilidades”. (Han, H., & Trimi, S. 2022). La mayoría de los modelos de aprendizaje automático, a pesar de lograr buenas soluciones técnicas a los problemas predictivos, terminan por no implementarse (Hassani et al., 2021) debido a desafíos prácticos, como problemas de integración, falta de datos adecuados, problemas de escalabilidad, o dificultades en la interpretación y aplicación de los resultados.

Es crucial tener en cuenta que la elección del algoritmo de segmentación debe ser cuidadosa teniendo en cuenta características como: problema a resolver, cantidad y tipo de datos necesarios, complejidad del problema, y capacidad para resolver un problema complejo.

En primer lugar, se debe evaluar la necesidad o el problema a resolver, existen algunos problemas que se pueden abordar mediante algoritmos específicos en el campo de la ciencia de datos.

Sí o No (A o B): Problemas que se pueden clasificar como "Sí" o "No" o con valores binarios (1 o 0), “este tipo de problemas se resuelven comúnmente utilizando algoritmos de clasificación” (Ranjani et al., 2022)

Variación o Excepción: Problemas que implican identificar patrones variados o encontrar la excepción, “estos problemas se pueden abordar mediante algoritmos de detección de excepciones u oddity detection” (Ranjani et al., 2022).

Cuánto o Cuántos: Problemas relacionados con mediciones, temperaturas, duraciones, valores numéricos, y cifras se resuelven utilizando algoritmos de regresión. “Los algoritmos de regresión analizan la relación entre variables, como la duración de un evento o la temperatura en una situación específica” (Ranjani et al., 2022).

Organización de Datos: Problemas relacionados con la gestión de datos en una organización se pueden resolver utilizando algoritmos de clustering. “Estos algoritmos organizan y agrupan conjuntos de datos basándose en características comunes, colores u otras características compartidas” (Ranjani et al., 2022).

Interpretabilidad: La interpretabilidad en la elección de modelos en ciencia de datos es fundamental, especialmente cuando el modelo o sus resultados tienen consecuencias importantes o se aplican a problemas complejos. Como indican Hassani et al. (2021), "la interpretabilidad es

crucial a menos que el modelo no tenga un impacto significativo o se relacione con un problema que haya sido bien estudiado." En contextos donde las decisiones pueden tener un gran impacto o donde el problema es intrínsecamente complejo, entender cómo funciona el modelo es esencial para garantizar que sus resultados sean fiables y útiles.

Por ejemplo, si un negocio minorista necesita comprender cómo ciertas tendencias a lo largo del tiempo "afectan la demanda futura, los métodos estadísticos son adecuados especialmente para los productos con demanda estable" (Hassani et al., 2021), ya que estos métodos permiten una interpretación clara de cómo las variables influyen en las predicciones, lo cual es importante para tomar decisiones informadas y estratégicas.

Sin embargo, Hassani et al. (2021) advierten que "confiar únicamente en la interpretabilidad (por ejemplo, utilizando modelos puramente estadísticos) puede proporcionarnos equidad, privacidad, confiabilidad, causalidad y confianza, a expensas de la precisión." Por lo tanto, al seleccionar un modelo, es crucial equilibrar la necesidad de interpretabilidad con la necesidad de precisión, considerando las implicaciones específicas del problema y el impacto de las decisiones basadas en el modelo.

Cantidad y tipos de datos: En las decisiones, es importante considerar el detalle de la información que se tiene, es diferente el presentar resultados que permitan tener una visión general sobre la reacción de los clientes a cierta campaña promocional que resultados que indiquen porque un cliente en específico tiene una probabilidad de compra baja o alta.

Tipo de resultado que entregara el modelo: El tipo de resultado se refiere al formato en el que se presenta el resultado al tomador de decisiones y varía según la elección del algoritmo que se implemente.

Hiperparámetros

Los hiperparametros son configuraciones ajustables de un modelo de aprendizaje automático que no se aprenden automáticamente del entrenamiento de datos. Los científicos de datos deben elegir valores adecuados para estos hiperparámetros para optimizar el rendimiento de su modelo.

Estrategias para la selección de hiperparámetros:

Experiencia previa y conocimiento del dominio: Utilizando la experiencia y conocimiento del dominio para seleccionar hiperparámetros.

Selección ad-hoc: Se elijen los valores de hiperparámetros de manera ad-hoc, a menudo utilizando valores predeterminados.

Selección rápida para una solución inicial: Se eligen rápidamente valores de hiperparámetros para tener una solución inicial y luego se planea como ajustar estos valores más tarde.

Evaluación Modelo

Otra consideración a tener en cuenta es como medir el desempeño de los modelos de predicción, esto se suele hacer por medio de métricas como el área bajo la curva ROC, la precisión, la medida H y métricas de evaluación operativa como el tiempo de cálculo, sin embargo, estás métricas suelen “suelen ser muy difíciles de interpretar para los responsables de la toma de decisiones empresariales y, por tanto, difíciles de ayudar a respaldar sus decisiones” (Coussement K., & Benoit, F., 2021). En este sentido es importante aplicar métricas más

interpretables en los negocios como el aumento del decil superior o el criterio de beneficio máximo.

Aplicación. Selección del Modelo: Al seleccionar un modelo, es fundamental considerar tanto la aplicación de la ciencia de datos como los objetivos de la aplicación. En este caso, el propósito final del proceso es tomar decisiones, por lo que es necesario utilizar un modelo altamente interpretable. Además, debe tenerse en cuenta el contexto en el que se toman estas decisiones. En el entorno minorista, que es altamente dinámico, las ventas pueden variar según múltiples factores.

Por ello, el interés principal radica en obtener una visión clara de cómo factores complejos, como promociones, días festivos o tendencias externas, están afectando las ventas, más que en capturar patrones estacionales o a largo plazo. Entendiendo que en este entorno el comportamiento de las ventas depende más de estas variables que de la secuencia temporal.

Para realizar predicciones de ventas en entornos minoristas, suelen emplearse modelos como LSTM y ARIMA, ya que son útiles para series temporales estacionales con patrones regulares y repetitivos en intervalos específicos. No obstante, ARIMA tiende a generar predicciones inexactas y requiere una cuidadosa selección de parámetros (p, d, q).

Con grandes volúmenes de datos y alta capacidad computacional, es posible lograr mayor precisión al incorporar una gama más compleja de variables. Un ejemplo es el uso de redes neuronales como LSTM, las cuales pueden identificar y recordar relaciones a largo plazo, siendo ampliamente utilizadas para pronósticos de series temporales. Sin embargo, LSTM es un modelo de caja negra, lo que dificulta explicar cómo o por qué se genera una predicción específica. Estas redes aprenden patrones a partir de datos secuenciales mediante capas no lineales, lo que complica la trazabilidad de sus decisiones.

De esta manera, los modelos considerados como la mejor elección fueron Random Forest Regressor, Decision Tree Regressor y LGBM Regressor, ya que se van a predecir valores continuos (ventas), las características (como las ventas de los últimos meses o promociones) son más relevantes que la secuencia temporal, y para tomar decisiones se requiere modelos más interpretables.

Estos modelos pueden aprovechar características adicionales (como promociones y festivos) para ofrecer una visión clara de los factores que están afectando las ventas. Sin embargo, dado que estos modelos no capturan la estacionalidad de los datos de manera inherente, es necesario transformar los datos añadiendo información temporal. Esto permite identificar patrones estacionales, es decir, si las ventas tienden a subir o bajar en ciertos meses o días del año.

Selección de características y variable independiente y objetivo: En esta etapa se seleccionaron las variables que tienen un impacto directo sobre las ventas futuras, como la cantidad de promociones, la familia de productos y los días festivos. Los modelos de árbol de decisión por si solos no tiene un mecanismo para manejar la dependencia temporal, en este caso se entrena con características adicionales que representan información temporal, como la media de ventas de los últimos 7 días, para ayudar al modelo a capturar patrones temporales. Estas características permiten que el árbol de decisión capture la variabilidad o las tendencias específicas de cada año y así tenga noción de la secuencia temporal para hacer las predicciones futuras.

Preparación de datos: Antes de entrenar el modelo, se ajustaron los datos para que pudieran ser procesados adecuadamente. Esto incluyó transformar las categorías (como la familia de productos) en valores numéricos y rellenar valores faltantes para evitar inconsistencias.

Además, se calculó la media de las ventas históricas para utilizarla como variable objetivo o referencia en la predicción.

Ajustes de hiperparámetros: Para la selección de los hiperparametros se utiliza la técnica GridSearchCV, que es una técnica de búsqueda en cuadrícula. Primero se definen valores que se quieren probar. Luego GridSearchCV evalúa cada combinación de hiperparámetros dentro del espacio que se definió, lo hace entrenando y validando el modelo para cada combinación de hiperparámetros utilizando validación cruzada. Después de probar todas las combinaciones de hiperparámetros, se selecciona la combinación que mejor rendimiento tiene según la métrica Mean Absolute Error (MAE).

Entrenamiento del modelo: Se seleccionaron tres modelo de regresión, basado en árboles de decisión (LightGBM), que es adecuado para manejar grandes cantidades de datos y múltiples variables. Este modelo se entrenó utilizando las características previamente seleccionadas para aprender las relaciones entre promociones, días festivos y ventas históricas, y así predecir las ventas futuras.

Figura 20

Evaluación Modelos

```
[LightGBM] [Info] Total Bins 1379
[LightGBM] [Info] Number of data points in the train set: 14771, number of used features: 9
[LightGBM] [Info] Start training from score 854.842777
Evaluación de modelos:
RandomForestRegressor - MAE: 64.59, MSE: 44362.10, RMSE: 210.62, R²: 0.99, MAPE: 18.29%
DecisionTreeRegressor - MAE: 86.00, MSE: 104194.70, RMSE: 322.79, R²: 0.97, MAPE: 23.21%
LGBMRegressor - MAE: 72.06, MSE: 41939.79, RMSE: 204.79, R²: 0.99, MAPE: 52.50%
```

Nota. Resultados de la Evaluación de los Modelos.

MAE (Mean Absolute Error): 64.59 - 86.00 - 72.06, este valor indica que, en promedio, las predicciones del modelo se desvían en aproximadamente 64.59 unidades de las ventas reales. Un MAE bajo sugiere que el modelo hace predicciones precisas.

MSE (Mean Squared Error): 44362.10 - 104194.70 - 41939.79, este valor penaliza más fuertemente los errores grandes en comparación con el MAE. Un MSE de 44362.10 indica que las diferencias entre las predicciones y los valores reales son, en general, relativamente pequeñas, aunque puede haber algunos errores significativos.

RMSE (Root Mean Squared Error): 210.62 - 322.79 - 204.79, este es el valor cuadrático medio de los errores y tiene las mismas unidades que la variable de salida (ventas). Un RMSE de 210.62 indica que, en promedio, el modelo se desvía en aproximadamente 210.62 unidades de las ventas reales. Es una métrica útil para entender la magnitud de los errores en las predicciones.

R^2 (R-squared): 0.99 - 0.97 - 0.99, este valor indica que el 99% de la variación en las ventas se puede explicar por las variables independientes del modelo. Un R^2 tan alto sugiere que el modelo es muy efectivo para capturar la relación entre las características y las ventas.

MAPE (Mean Absolute Percentage Error): 18.29% - 23.21% - 52.50%, este valor indica que, en promedio, el modelo tiene un error del 18.29% en sus predicciones. Aunque es relativamente alto, un MAPE por debajo del 20% se considera generalmente aceptable en muchos contextos.

Definir KPI

Al tomar una decisión se debe buscar que esta sea tangible y específica, estableciendo indicadores de rendimiento. “No importa si estas medidas se etiquetan como KPIs o KRIs (Key Results Indicators), en su esencia son un conjunto de métricas específicas que cada negocio

utiliza para cuantificar el logro de aquellos objetivos que reflejan el desempeño o resultados de la organización en su conjunto para un determinado período de tiempo.” (Marín, G. J., et al. 2020).

Los KPIs son herramientas esenciales para medir y evaluar datos y hechos relevantes en relación con los objetivos de negocio. Según Balboni et al. (2013), “estos KPIs deben estar asociados con el objetivo de negocio y justificarse con los resultados previstos, tangibles e intangibles, que se espera obtener”.

Si no se han establecido los KPI, es difícil distinguir datos irrelevantes y aquellos “datos inteligentes que brindan información relevante, precisa y oportuna para cada KPI de un negocio” (Marín, G. J., et al. 2020), lo que puede resultar en un análisis poco efectivo o confuso.

Al analizar los datos en función de estos KPIs, se puede obtener una visión clara del rendimiento de los modelos y si los resultados obtenidos cumplen con los objetivos. Así, la “medición de los indicadores clave de rendimiento (KPI) identificados durante las etapas de ideación o inicio avalan los beneficios del modelo”. Karimi Dastgerdi, A., & Javdani Gandomani, T. (2021) lo que permite a los responsables de tomar decisiones evaluar qué tan bien se están logrando los objetivos, si el modelo está generando el valor esperado y dónde podrían ser necesarios ajustes.

Finalmente, una vez definidos los KPI, “es la recopilación, almacenamiento, procesamiento y análisis de datos relacionados con KPI utilizando herramientas de inteligencia minorista adecuadas de lo que más debemos preocuparnos.” (Marín, G. J., et al. 2020).

Aplicación. Para definir KPIs efectivos, primero debemos preguntarnos: ¿Cuál es la meta que se quiere alcanzar? Es crucial establecer esta meta en valores numéricos concretos para que el progreso pueda ser medido de manera objetiva. Además, es necesario definir los criterios que la solución debe cumplir, que deben responder a preguntas clave como:

- ¿Cómo validaremos los resultados obtenidos?
- ¿Cómo controlaremos el progreso de la ejecución a través de indicadores líderes?
- ¿Cuándo planeamos lograr estos resultados, y cuáles son los valores objetivo?

Figura 21

Crecimiento de Ventas Año 2024

month_name	sales_2023	sales_predicted_2024_RandomForestRegressor	sales_predicted_2024_increase_promotion_RandomForestRegressor	growth_2023	growth_predicted_2024
January	313698.498204	297056.639411	294898.812992	0.000000	-5.305049
February	288534.509892	276085.125138	280753.856079	-8.021711	-4.314695
March	313689.789778	307812.674094	308957.862726	8.718292	-1.873544
April	350701.565039	300482.023371	300419.581581	11.798846	-14.319737
May	312273.395204	309791.344860	310012.156084	-10.957513	-0.794832
June	307671.206879	301367.442028	311812.079976	-1.473769	-2.048864
July	306983.169067	320662.062582	328324.242714	-0.223628	4.455910
August	298315.684973	303872.369878	314449.976254	-2.823440	1.862686
September	301010.831889	314877.032728	320976.562173	0.903455	4.606545
October	324185.842051	330683.434835	335258.431578	7.699062	2.004280
November	312168.995974	324965.421110	335482.697790	-3.706777	4.099198
December	364259.232032	371651.104682	378460.709167	16.686550	2.029289

Fuente. Crecimiento de Ventas Predicho para el Año 2024

Primer objetivo Mensual: Se usa el modelo de predicción para estimar las ventas del cada mes del año 2024. Estas predicciones sirven como base para los objetivos de ventas así:

Para ajustar los objetivos de ventas, utilizamos tanto la tasa de crecimiento histórica como la tasa de crecimiento predicha. En este caso, el modelo predice que en septiembre de 2024, las ventas aumentarán un 4.606545% en comparación con las ventas de septiembre de 2023. Si establecemos este aumento como nuestro objetivo, y teniendo en cuenta que las ventas para este mes el año anterior fueron de 301,010.83, las ventas para septiembre de 2024 deberían ser de aproximadamente 315,724.25 unidades.

Si comparamos este valor con las ventas estimadas al considerar un incremento del 10% en las promociones, nuestras ventas alcanzarían alrededor de 320,976.56 unidades. Esto significa que, con el aumento previsto del 10% por las promociones, estaríamos logrando la meta de ventas para 2024.

Segundo objetivo trimestral: Basándote en tus predicciones mensuales, se crea una proyección para todo el año con el fin de fijar objetivos de ventas trimestrales. Se suman las predicciones de cada mes y establecer un objetivo para fin de año.

Enero: 297,056 unidades

Febrero: 276,085 unidades

Marzo: 307,812 unidades

El objetivo del primer trimestre sería alcanzar ventas de 880,953 unidades vendidas.

Plan de Acción

La información se transforma en conocimiento cuando los encargados de tomar decisiones combinan su experiencia con los datos presentados y los resultados obtenidos. Este proceso implica “la interpretación de la información percibida, creando el conocimiento que permitirá determinar las alternativas de decisión y la posterior elección de la mejor” (Rodríguez-Cruz, Y., & Pinto, M., 2018).

El plan de acción permite identificar o crear posibles soluciones, para enfrentar la situación o problema. En esta fase, el análisis de la información es crucial, ya que facilita la asimilación y comprensión de todos los datos recopilados, lo que a su vez contribuye a la creación de conocimiento (Rodríguez-Cruz & Pinto, 2018). Sin embargo, los encargados de tomar decisiones pueden interpretar los resultados de diversas maneras, incluyendo estar en desacuerdo con ellos, malinterpretarlos o incluso ignorarlos. Como señala Solano-Brenes (2013),

"únicamente el tomador de decisiones, por medio del buen juicio, experiencia, inteligencia, educación, tiempo disponible, etc., sabrá en qué momento dejar de recolectar información y decidir lo que crea más conveniente."

En esta etapa es fundamental “buscar caminos alternativos para alcanzar una meta. Estas soluciones van desde las que ya se tienen hasta las que se diseñan a la medida. Cuando quienes toman las decisiones buscan soluciones probadas, utilizan ideas que se han puesto en marcha o siguen un benchmarking al considerar experiencias similares de empresas competidoras o líderes en su área de influencia. Por otra parte, cuando las soluciones son a la medida, es necesaria la combinación de nuevas ideas para lograr que la solución sea específica al requerimiento”. (Franklin Fincowsky, E. B., 2011).

Aplicación. En esta fase se exploran diferentes alternativas y se visualizan los resultados en forma de gráficos interactivos. Al implementar modelos como RandomForestRegressor, DecisionTreeRegressor y LGBMRegressor, se analiza cómo diversas estrategias, como el aumento del 10% en las promociones en distintos días, pueden afectar las ventas predichas. Además, se evalúa el impacto de estas promociones sobre cada familia de productos en particular. Se muestran cuáles serán las familias de productos más vendidas cada mes del año, en que días se tendrán más ventas, que meses se tendrán la mayor cantidad de ventas de cada familia de productos.

Esto permite visualizar cómo diferentes acciones podrían influir en el desempeño futuro, generando múltiples alternativas de decisión. De esta manera, es posible identificar no solo las ventas estimadas para cada mes, sino también determinar en qué mes se esperan las mayores ventas para cada familia de productos. Disponer de este tipo de información será de gran ayuda para la persona encargada de mantener el almacén ya que permite conocer en qué momento se

puede agotar un producto es vital para realizar un pedido a los proveedores adecuadamente. Esto evita que los clientes intenten comprar un artículo agotado y que los trabajadores tengan problemas para ubicar las unidades sobrantes en el almacén.

De esta etapa se pueden generar diferentes alternativas como, aumentar los precios en diciembre ya que se predice un aumento en la demanda, buscando maximizar los márgenes de ganancia. Por otro lado, para meses como febrero y septiembre que se proyectan bajas ventas se pueden ofrecer idear estrategias de descuentos, reducir inventario y diseñar campañas publicitarias para atraer clientes.

Figura 22

Predicción de Ventas Mensuales Año 2024

Nota. Predicción de ventas mensuales para el Año 2024 Con y Sin Aumento de Promociones.

Priorizar Decisiones

La etapa de priorización de decisiones es crucial para asegurar que la opción seleccionada sea la más adecuada para alcanzar los objetivos propuestos. Luego de identificar todas las

alternativas, “el tomador de decisiones evalúa de manera crítica cada una; considerando cuidadosamente las ventajas y desventajas de cada alternativa” (Solano-Brenes, A. I., 2013). La evaluación de las diferentes alternativas generadas en la etapa anterior se realiza “teniendo en cuenta la concordancia con los objetivos de la empresa y los recursos.

Además, la alternativa elegida debe ser factible y contribuir a la resolución del problema. Hay que tener en cuenta los posibles problemas futuro y las consecuencias asociadas a cada una de las alternativas” (Canós , L., et al., 2012). Si se identifican muchas soluciones con facilidad, es posible aumentar los criterios de selección para reducir el número de candidatos como:

- Impacto de la decisión
- Riesgos potenciales
- Tiempo de ejecución.
- Lecciones aprendidas de decisiones pasadas.

Tomar una decisión implica escoger entre varias alternativas la que menor riesgo conlleve en llegar a la meta. “Este paso hace hincapié en determinar los resultados que se esperan y el costo relativo de cada alternativa”. (Franklin Fincowsky, E. B., 2011), el objetivo es encontrar una solución que sea satisfactoria, priorizando su viabilidad y efectividad más que la optimización absoluta. Es fundamental recordar que enfocarse únicamente en la maximización de utilidades puede ser peligroso, ya que este es un concepto a corto plazo. Por ejemplo, una decisión de eliminar líneas de productos debe considerar más allá de la utilidad que estos generan. “Para este efecto, son importantes los conceptos de maximizar, satisfacer y optimizar. Maximizar es tomar la mejor decisión posible con el mayor beneficio al menor costo y el mayor rendimiento esperado. Satisfacer significa que en la búsqueda de alternativas se elige la primera

aceptable o adecuada de acuerdo con el criterio o meta definidos. Optimizar significa alcanzar el mejor equilibrio entre metas múltiples". (Franklin Fincowsky, E. B., 2011).

"Los responsables de la toma de decisiones tienen que considerar distintos tipos de consecuencias. Pueden intentar predecir los efectos en el comportamiento financiero o de gestión de la empresa. Es de esperarse que no va a ser posible predecir los resultados con toda precisión, pero pueden servir para prepararse para un futuro incierto y sus consecuencias potenciales y generar planes de contingencia". (Franklin Fincowsky, E. B., 2011).

Además, para abordar indecisiones, es importante reflexionar, hacer preguntas concretas y aceptar la posibilidad de equivocarse, ya que esto forma parte del proceso de toma de decisiones. "No debe pensarse que existe un conjunto maravilloso de fórmulas que una vez que se aprenden proporcionarán respuestas gloriosas a todos los problemas. No hay tal. Todavía se necesitan el juicio, la experiencia, la intuición y el coraje humanos para administrar una empresa." (Gallagher, C. A., & Watson, H. J., 1988).

También es posible incorporar diversas perspectivas del problema, como las recomendaciones de expertos en administración. Sin embargo, es esencial siempre tener en cuenta el contexto de la decisión, evaluando tanto los factores controlables como los incontrolables, así como los posibles resultados de cada alternativa.

Aplicación. Para tomar la decisión se deben tener en cuenta todas las cifras relacionadas con el valor que consideramos, elementos que reducen ese valor o lo afectan. La toma de decisiones basada en datos puede realizarse alrededor de algunas cifras generales, pero no de una cifra única. Utilizar todos los datos relevantes permite una verdadera toma de decisiones basada en datos.

En este caso las predicciones permiten evaluar diferentes escenarios y alternativas para predecir las ventas futuras utilizando varios modelos de predicción, como RandomForestRegressor, DecisionTreeRegressor, y LGBMRegressor. Además, contempla el análisis de escenarios hipotéticos, como el aumento del 10% en las promociones en diferentes momentos del año y para cada grupo de productos, permitiendo analizar las consecuencias potenciales de esa decisión sobre las ventas futuras.

Este enfoque facilita la comparación entre las alternativas, asegurando que la opción seleccionada sea la más adecuada para cumplir con los objetivos establecidos, considerando tanto las oportunidades como posibles pérdidas y caídas de ventas.

Con estos resultados no se intenta predecir el futuro al 100%, porque este tipo de análisis es probabilístico, sí pronostican qué podría suceder. Así se entienden las correlaciones entre variables y cómo podrían comportarse en un futuro. La aplicación de ciencia de datos ofrece resultados sobre las ventas predichas para cada mes del año sin aumento en la cantidad de promociones y con aumento.

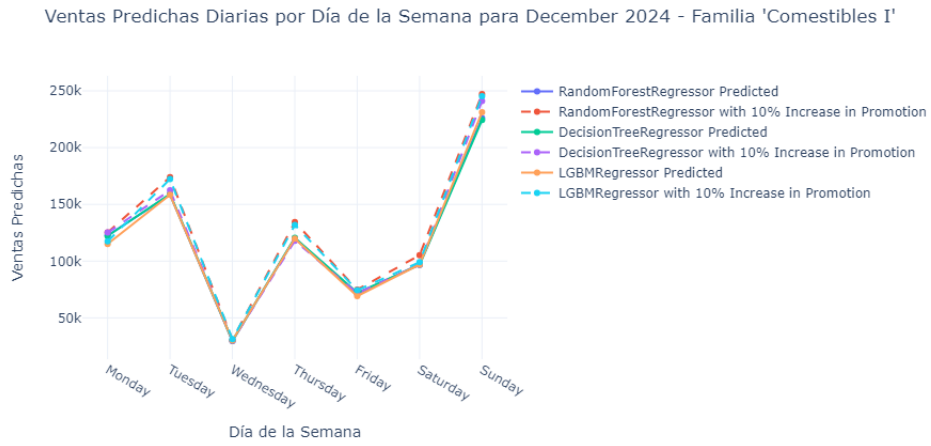
Figura 23

Ventas Mensuales Predichas para la Familia de Productos Comestibles

Nota. Ventas mensuales predichas para la familia de productos Comestibles para cada mes del año 2024 con y Sin Aumento de Promociones.

Figura 24

Ventas Diarias Predichas para la Familia de Productos Comestibles



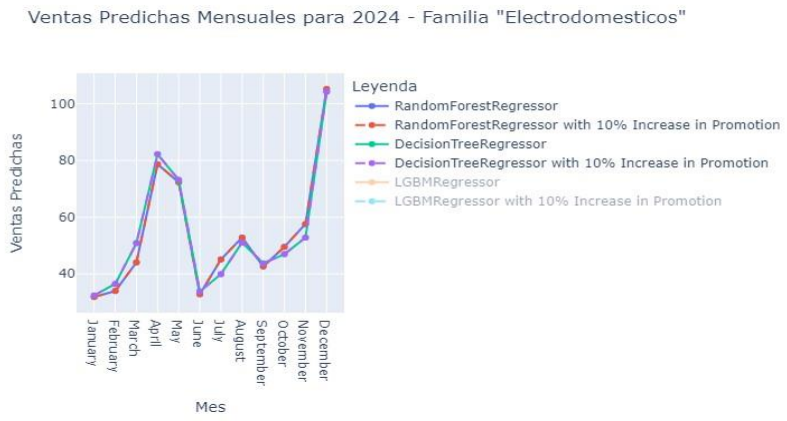
Nota. Ventas diarias predichas para la familia de productos Comestibles para el mes de diciembre del año 2024 con y Sin Aumento de Promociones.

Tenemos resultados de las ventas predichas para cada día de la semana, permitiendo seleccionar el mes y la familia de productos que se desea analizar. De esta manera se debe tomar decisiones para asegurar que los domingos del mes de diciembre haya constante reabastecimiento de comestibles ya que se proyectan picos de ventas. Por otro lado, los miércoles de diciembre en los que se predice una caída en la demanda se debe optar por campañas de descuentos, considerando que el aumento de la cantidad de productos en promoción no es una alternativa válida ya que no resulta en un incremento significativo de ventas según las predicciones. Otra posible decisión podría ser la de eliminar la venta de productos como artículos como libros ya que las ventas son bajas y solo venderlos en meses específicos, en cambio optar

por aumentar el espacio para vender electrodomésticos ya que la proyección de aumentos en ventas es mayor.

Figura 25

Ventas Mensuales Predichas para la Familia de Productos "Electrodomésticos"



Nota. Ventas mensuales predichas para la familia de productos Electrodomésticos para cada mes del año 2024 con y Sin Aumento de Promociones.

Figura 26

Ventas Mensuales Predichas para la Familia de productos "Libros"

Nota. Ventas mensuales predichas para la familia de productos Electrodomésticos para cada mes del año 2024 con y Sin Aumento de Promociones.

Ejecutar

"Una vez identificada la mejor alternativa de decisión se realiza la etapa de implementación y control de la decisión que permite establecer los cursos de acción para implementar la decisión y en el mismo se analiza si la misma da solución a la situación-problema a partir de los resultados obtenidos" (Rodríguez-Cruz, Y., & Pinto, M., 2018).

Esto puede implicar la asignación de recursos, la asignación de tareas y responsabilidades, y la puesta en marcha de acciones concretas para llevar a cabo la decisión tomada. Para una ejecución efectiva, es esencial desarrollar un plan de acción que detalle los siguientes aspectos:

- Qué se va a hacer: Definir claramente las actividades y pasos necesarios para implementar la decisión. "Se desarrollan las acciones que conlleva la alternativa elegida para solucionar el problema" (Canós, L., et al., 2012)

- Quién lo va a hacer: Asignar tareas y responsabilidades específicas a las personas o equipos encargados de la implementación. "Una decisión técnicamente correcta debe ser aceptada y apoyada por las personas que se encargarán de su implementación para que haya una actuación efectiva basada en la decisión". (Franklin Fincowsky, E. B., 2011).

- Qué tiempo de duración tiene la actividad: Establecer un cronograma con plazos concretos para cada actividad. "Quienes implementan la decisión deben comprender la elección y los factores que mediaron para tomarla, asumir y mantener el compromiso de ejecutarla, ordenar en forma cronológica los pasos para que sea operativa asignando los recursos necesarios y calcular los tiempos consecuentes para culminarla" (Franklin Fincowsky, E. B., 2011).

- En qué lugar se ejecutará: Determinar el lugar o contexto donde se llevarán a cabo las actividades.

- Como se controlará: En el entorno minorista, donde las condiciones pueden cambiar rápidamente, el seguimiento es vital para asegurar que la decisión sigue siendo válida y efectiva. Este seguimiento implica construir preguntas que permitan evaluar los resultados y determinar si la implementación de la solución seleccionada está logrando la meta deseada, ya que "la implementación de la solución seleccionada no logrará de forma automática la meta deseada" (Franklin Fincowsky, 2011).

Esta etapa es fundamental para “comprobar si la puesta en marcha de la decisión es la más adecuada y si se alcanzan los resultados deseados”. (Canós, L., et al., 2012). El seguimiento puede revelar que la decisión tomada necesita ajustarse o que las metas deben revisarse, “se realiza un control evaluando las acciones pasadas y si algo no es correcto, se reinicia el proceso”. (Canós, L., et al., 2012).

Aplicación. Los resultados de las predicciones ofrecen conocimiento a futuro sobre la tendencia de venta. Acceder a esta información permiten realizar plan para lograr que la decisión sea tangible y alcanzable, además, realizar un seguimiento efectivo de la decisión tomada para verificar si se están alcanzando los objetivos:

Evaluación de resultados: Predecir las ventas antes, durante y después de la implementación de la decisión para determinar su efectividad, verificar si se solucionó el problema identificado en los tiempos estimados de no ser así revisar los resultados de las fases, corregir y en caso de ser necesario plantear una nueva decisión.

Reevaluación de objetivos: Los resultados permiten revisar si los objetivos establecidos son realistas y alcanzables, permitiendo ser ajustados si es necesario en función de las ventas predichas.

Análisis temporal: Desglosar las cantidades de venta en intervalos de tiempo específicos (mensual, diario) permite observar tendencias y patrones a lo largo del tiempo.

Revisar la estrategia: Si los resultados no cumplen con las expectativas, se debe evaluar en qué momento se logrará el objetivo propuesto y qué decisiones pueden acercar a la organización a ese objetivo.

Análisis de desviaciones: Identificar y analizar las razones detrás de las diferencias entre los resultados esperados y los reales, considerando factores como la ejecución de campañas promocionales, la disponibilidad de productos y la efectividad de la comunicación.

Control de la Decisión: Al utilizar métricas para evaluar los modelos, se puede determinar la precisión de las ventas predichas, lo que nos permite estimar con cierta probabilidad la diferencia entre los valores predichos y los posibles valores reales.

Conclusiones

La ciencia de datos transforma los datos en conocimiento valioso, ofreciendo información clave, automatizando decisiones y prediciendo resultados. Su aplicación abarca una amplia variedad de enfoques en la toma de decisiones, ya sea como herramienta de guía, apoyo o para la automatización de procesos. Cuando se utiliza como guía, proporciona un marco analítico para abordar problemas, utilizando modelos descriptivos, predictivos y prescriptivos. Como apoyo, la ciencia de datos ofrece información crítica para la toma de decisiones, aun cuando no exista un modelo definitivo para una solución. Finalmente, en la automatización, la ciencia de datos permite crear modelos que pueden integrarse en sistemas informáticos para tomar decisiones automáticamente.

En la implementación de ciencia de datos en los negocios Retail, los principales desafíos son la escasez de datos, las dificultades relacionadas con la recopilación, el almacenamiento, el uso, el análisis, la privacidad y la confianza en los datos, la mala gestión de datos, la idea de que las soluciones tecnológicas resultan difíciles de implementar y costosas, además, factores relacionados con el entorno altamente cambiante en el que operan: alta competencia, nuevas tecnologías, la presencia de nuevos competidores, cambios en la legislación o disturbios políticos, entre otros. La interacción de estos factores hace que la previsión precisa sea un desafío, al ser impredecible y fluctuante, complicando la creación de modelos de ciencia de datos robustos.

La ciencia de datos permite dar sentido y valor a los datos para la toma de decisiones en negocios Retail, al permitir aprovechar la creciente disponibilidad de datos para transformarlos en conocimiento, utilizando enfoques basados en datos y aprovechando las herramientas avanzadas de datos para obtener resultados más precisos y rápidos en un entorno altamente

cambiante. Tiene el potencial de generar beneficios significativos en el sector minorista, abarcando todas las etapas de la cadena de valor. Desde la planificación, donde permite pronosticar ventas y medir indicadores clave de desempeño, hasta la optimización de la cadena de suministro, facilitando la gestión de inventarios y el rendimiento de los proveedores. Técnicas como el análisis predictivo y el aprendizaje automático pueden brindar a estos negocios la oportunidad de identificar patrones de comportamiento y mejorar la gestión en áreas críticas como inventario y personalización de la experiencia del cliente.

Tomar una decisión basada en datos no es solamente elegir la tecnología de análisis adecuada para identificar la próxima oportunidad estratégica sino debe ser un enfoque integral que permita entender el negocio, sus objetivos, expectativas, sus reglas, etc. Algoritmos sencillos creados sobre situaciones ideales, pero no enfocadas en los negocios dan resultados inútiles.

La aplicación y elección de un modelo de ciencia de datos en la toma de decisiones depende de varios aspectos como el tamaño del conjunto de datos, los objetivos del negocio y del análisis. Si se busca comprender las relaciones subyacentes, los métodos estadísticos suelen ser más adecuados. Por otro lado, cuando se prioriza la precisión de las predicciones, se opta por técnicas de aprendizaje automático o una combinación de ambos enfoques. Los modelos de aprendizaje automático no permiten una buena interpretación de los resultados, por lo tanto, en el futuro, es esencial incorporar métodos de inteligencia artificial explicable (XAI), con el objetivo de que los modelos no solo ofrezcan una respuesta, sino que expliquen cómo y por qué se tomó una decisión. Esto mejorará la confianza en las decisiones impulsadas por IA, ya que la transparencia, imparcialidad y precisión son factores clave al tomar una decisión. Esto es un todo reto para la inteligencia artificial ya que todas estas explicaciones aún tienen limitaciones para explicar decisiones complicadas.

Independientemente del enfoque que se dé, el proceso de aplicación de ciencia de datos para la toma de decisiones debe estar alineado con los objetivos comerciales del negocio y apoyarse en metodologías sólidas como CRISP-DM, que guían desde la comprensión del negocio hasta la evaluación de los resultados. La implementación de estas metodologías permite no solo una toma de decisiones más fundamentada, sino también un enfoque integral que mejora la eficiencia operativa y la rentabilidad.

El proceso de analítica debe estar guiado por un profundo entendimiento del negocio, comenzando por identificar, definir y diagnosticar qué está ocurriendo, cuáles son los objetivos, los desafíos que enfrenta, los recursos disponibles, y los datos que se tienen. Es crucial comprender cómo están estructurados esos datos, cómo se almacenan, y qué tan limpios y correctos son para luego transformarlos en un formato adecuado libre de errores y vacíos.

Una vez los datos están preparados, se debe analizar cuáles son las variables más relevantes, cómo se relacionan entre sí y qué factores influyen en sus cambios. Con este conocimiento, es posible formular preguntas clave sobre los datos, agruparlos para capturar su esencia, y generar predicciones o proyecciones que permitan tomar decisiones informadas.

Finalmente, para tomar estas decisiones se requiere generar diversas alternativas de decisión, y evaluarlas para seleccionar la mejor opción y finalmente ejecutarla teniendo en cuenta que es esencial mantener un ciclo de mejora continua, monitoreando los resultados, ajustando los objetivos según el contexto y formulando nuevas preguntas que puedan generar nuevas decisiones y enfoques estratégicos.

En el proceso de toma de decisiones, las predicciones generadas por la ciencia de datos nos permiten formular preguntas, plantear hipótesis y validarlas utilizando diferentes técnicas, según el problema a resolver. En un contexto donde diversos atributos interactúan entre sí, es

donde la ciencia de datos brinda beneficios reales ya que los algoritmos pueden aprender de estos e identificar uno o más atributos derivados que brindan información de un problema. Siendo de especial importancia para la toma de decisiones ya que, para cada alternativa de decisión, la ciencia de datos facilita la proyección de resultados y consecuencias, lo que ayuda a seleccionar la mejor opción para alcanzar los objetivos propuestos. Además, a partir del aprendizaje de las variables permite realizar predicciones para establecer indicadores clave de rendimiento (KPI) que controlen las decisiones, diseñar cronogramas con plazos concretos para cada actividad y anticiparnos a posibles inconvenientes como pérdidas, desabastecimiento o caída en ventas.

Recomendaciones

Para una futura aplicación en el sector Retail colombiano donde existe desconfianza hacia el análisis de datos y la percepción de que los modelos de analítica no aportan beneficios claros, se recomienda elegir un problema clave que sea crítico para el negocio, donde los resultados sean fácilmente medibles en términos de incremento de ingresos, ahorros, reducción de tiempos, optimización de recursos, y la mejora en la satisfacción del cliente. No basta con brindar resultados precisos sino proporcionar una solución que acerque al negocio a sus objetivo y genere beneficios visibles y veloces para ganar así confianza. Tambien, es importante que se considere la privacidad de los datos, evitar extraer o pedir datos con información personal de los clientes o empleados, explicar claramente como se gestionaran los datos y para que se usarán, demostrando compromiso con la seguridad de los datos puesto que un proyecto donde no se considere la protección de datos puede generar desconfianza en el negocio, además de ser propenso a problemas legales.

Por otro lado, se requiere que los negocios empleen sistemas de información bien diseñados e implementados que garanticen información adecuada y que los datos recopilados sean relevantes y de alta calidad (fiables, completos, concisos, coherentes, válidos y precisos). Un modelo de IA es tan bueno como la estrategia de datos que lo respalda.

Finalmente, el juicio humano no debe dejarse a un lado al tomar una decisión. Por lo que además de información precisa, es esencial integrar los conocimientos derivados del análisis de la experiencia humana, al tomar una decisión se debe hacer uso de la creatividad, se debe tener experiencia, se debe contar con conocimiento profundo del problema, de la organización y su entorno y si es posible también incorporar diversas perspectivas del problema, como las recomendaciones de expertos en administración.

Referencias Bibliográficas

Abad-Segura, E., & González-Zamar, M.-D. (2020). Global Research Trends in Financial Transactions. Mathematics, 8(4), 614

Abad-Segura, E., González-Zamar, M. D., & López-Meneses, E. (2022). El proceso de toma de decisiones basado en métodos cuantitativos: análisis de tendencias en el ámbito corporativo. Revista De Métodos Cuantitativos Para La Economía Y La Empresa, 34, 118–136. https://doi.org/10.46661/revmetodoscuanteconempresa.5135

Abbas, W., Usman, M., & Qamar, U. (2022). Churn prediction of customers in a retail business using exploratory data analysis. In 2022 International Conference on Frontiers of Information Technology (FIT) (pp. 130-135). IEEE. https://doi.org/10.1109/FIT57066.2022.00033

Aburto, Luis, and Richard Weber. (2003) "Demand forecast in a supermarket using a hybrid intelligent system." In Design and application of hybrid intelligent systems: 1076–1083.

Aburto, Luis, and Richard Weber. (2007) "Improved supply chain management based on hybrid demand forecasts." Applied Soft Computing 7 (1): 136–44.

Alexander, D.T., & Lyytinen, K.J. (2017). Organizing Successfully for Big Data to Transform Organizations. Americas Conference on Information Systems.

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. Journal of Retailing and Consumer Services, 8(3), 147-156. https://doi.org/10.1016/S0969-6989(00)00011-4

Alurkar, A. A., Ranade, S. B., Joshi, S. V., Ranade, S. S., Sonewar, P. A., Mahalle, P. N., & Deshpande, A. V. (2017). A proposed data science approach for email spam classification

using machine learning techniques. In 2017 Internet of Things Business Models, Users, and Networks (pp. 1-5). https://doi.org/10.1109/CTTE.2017.8260935

Al-Zahrani, A., & Al-Hebbi, M. (2022). Big Data Major Security Issues: Challenges and Defense Strategies. Tehnički glasnik, 16(2), 197-204. https://doi.org/10.31803/tg-20220124135330

Arriagada Benítez, Mauricio. (2020). Ciencia de Datos: hacia la automatización de las decisiones. Ingeniare. Revista chilena de ingeniería, 28(4), 556-557. https://dx.doi.org/10.4067/S0718-33052020000400556

Arunraj, N., Ahrens, D., & Fernandes, M. (2016). Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry. International Journal of Operations Research and Information Systems, 7, 1-21. https://doi.org/10.4018/IJORIS.2016040101.

Aversa, J., Hernandez, T., & Doherty, S. (2021). Incorporating big data within retail organizations: A case study approach. Journal of Retailing and Consumer Services, 60, 102447. doi: 10.1016/j.jretconser.2021.102447

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the International Conference on Data Mining (IADS-DM)

Bala, P.K. (2010). Decision tree based demand forecasts for improving inventory performance. 2010 IEEE International Conference on Industrial Engineering and Engineering Management, 1926-1930.

Balboni, F., Finch, G., Rodenbeck Reese, C., & Shockley, R. (2013). Analítica de datos: un proyecto de generación de valor. Cómo transformar Big Data en resultados, a través de la analítica. IBM Global Business Services, IBM Institute for Business Value. https://atenea.epn.edu.ec/handle/25000/311

Bellini, P., Palesi, L.A.I., Nesi, P. (2023). Multi Clustering Recommendation System for Fashion Retail. Multimed Tools Appl 82, 9989–10016. https://doi.org/10.1007/s11042-021-11837-5

Bettis-Outland, H. (2012). Decision-making's impact on organizational learning and information overload. Journal of Business Research, 65(6), 814-820. https://doi.org/10.1016/j.jbusres.2010.12.021

Biron, K., Mansoor, W., Miniaoui, S., Atalla, S., Mukhtar, H., & Bin Hashim, K. F. (2019). Data Science Tools for Crime Investigation, Archival, and Analysis. In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1263-1266). doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00235

Bocangel Carbajal, J. L., et al. (2020). Análisis de las características que identifica a un usuario de practisis premium: variables que deciden para convertirse de una cuenta freemium a premium. Retrieved from http://hdl.handle.net/10757/655935.

Bratina,D. & Faganel,A. (2008). Forecasting the Primary Demand for a Beer Brand Using Time Series Analysis. Organizacija,41(3) 116-124. https://doi.org/10.2478/v10051-008-0013-7

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cabeza de Vergara, L., & Muñoz Santiago, A. E. (2010). Análisis del proceso de toma de decisiones, visión desde la PYME y la gran empresa de Barranquilla. Cuadernos Latinoamericanos de Administración, VI (10), 9-40.

Cam Gensollen, C. R. (2022). Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa. Ing. ind. (Lima), pp. 189-216.

Canós Darós, L., Pons Morera, C., Valero Herrero, M., & Maheut, J. P. D. (2012). Toma de decisiones en la empresa: Proceso y clasificación. Universitat Politècnica de València. https://doi.org/10.4995/learning_objects_0000

Chan, H., & Wahab, M. I. M. (2024). A machine learning framework for predicting weather impact on retail sales. Supply Chain Analytics, 5, 100058. https://doi.org/10.1016/j.sca.2024.100058

Chávez García, E. M., Arguello Pazmiño, A. M., Viscarra Armijos, C. P., Aro Sosa, G. L., & Albarrasín Reinoso, M. V. (2018). Inteligencia Artificial en la toma de decisiones gerenciales. Dilemas Contemporáneos: Educación, Política y Valores, (Número: Edición Especial), Artículo no. 41. https://dilemascontemporaneoseducacionpoliticayvalores.com/index.php/dilemas/article/view/630/825

Chávez Larios, J. A., & Saucedo Martínez, N. (2018). Aplicación teórica de un modelo de análisis predictivo para desarrollar estrategias competitivas en MiPYMES. Repositorio De La Red Internacional De Investigadores En Competitividad, 10(1). https://riico.net/index.php/riico/article/view/1398

Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018). Intelligent Sales Prediction Using Machine Learning Techniques. 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 53-58.

Chern, C. C., Wei, C. P., Shen, F. Y., & Fan, Y. N. (2015). A sales forecasting model for consumer products based on the influence of online word-of-mouth. Information Systems and e-Business Management, 13(3), 445–473.

Chiang, L.-L. (Luke), & Yang, C.-S. (2018). Does country-of-origin brand personality generate retail customer lifetime value? A Big Data analytics approach. Technological Forecasting and Social Change, 130, 177-187. https://doi.org/10.1016/j.techfore.2017.06.034

Chowdhury, M. T. A., & Sharma, N. (2021). Citizenly: A platform to encourage data-driven decision making for the community by the community. In 2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics) (pp. 359-364). Melbourne, Australia. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics53846.2021.00064

Chu, C.-W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics, 86(3), 217-231. https://doi.org/10.1016/S0925-5273(03)00068-9

Claudino, J. G., Capanema, D. de O., de Souza, T. V., Serrão, J. C., Machado Pereira A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. Sports Medicine - Open, 5(1), 28. https://doi.org/10.1186/s40798-019-0202-3

Cortiñas Ugalde, M., Chocarro Eguaras, R., & Villanueva, M. L. (2010). La heterogeneidad de los consumidores en la valoración de la gestión minorista. Revista Española de Investigación de Marketing, 14(1), 91-114.

Coussement, K., & Benoit, F. (2021). Interpretable data science for decision making. Decision Support Systems, 113664. https://doi.org/10.1016/j.dss.2021.113664

Cruz, C. (2017). Impacto de los minimercados en el retail colombiano. Retrieved from https://repository.unimilitar.edu.co/bitstream/handle/10654/14428/CruzCarlosEduardo2016.pdf?sequence=1&isAllowed=y.

DANE. (2019). Muestra Mensual de Comercio al por Menor -MMCM-. Departamento Administrativo Nacional de Estadística. https://www.dane.gov.co/files/investigaciones/fichas/comercio_servicios/ficha_mmcm.pdf

DANE. (2023). Boletín Técnico Gran Encuesta Integrada de Hogares (GEIH), Principales indicadores del mercado laboral. https://www.dane.gov.co/files/investigaciones/boletines/ech/ech/bol_empleo_mar_23.pdf

DANE. (2024). PIB Información técnica I semestre 2024. https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-trimestrales/pib-informacion-tecnica

Dellino, G., Laudadio, T., Mari, R., Mastronardi, N., & Meloni, C. (2015). Sales Forecasting Models in the Fresh Food Supply Chain. In Proceedings of the International Conference on Operations Research and Enterprise Systems (ICORES-2015) (pp. 419-426). ISBN: 978-989-758-075-8. DOI: 10.5220/0005293204190426

Dev, M., Kumar, A., Kumar, G., & Singh, G. (2022). Data science related to big data, data-driven decision making and its application. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 221-227). IEEE. https://doi.org/10.1109/ICTACS56270.2022.9988578

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73. https://doi.org/10.1145/2500499

Díaz Parra, J. S. & Arango Moreno, J. F. (2020). Análisis bibliométrico de la producción científica sobre la aplicación de las ciencias de datos para la toma de decisiones en análisis de riesgos, 2015-2020. http://hdl.handle.net/11349/29244.

Ebadi, A., Gauthier, Y., Tremblay, S., & Paul, P. (2019). How can Automated Machine Learning Help Business Data Science Teams? In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 1186-1191). DOI: 10.1109/ICMLA.2019.00196

Elmasdotter, A., & Nyströmer, C. (2018). A comparative study between LSTM and ARIMA for sales forecasting in retail.

Escobar Gutiérrez, E., Ramírez Roa, D. P., Quevedo Hernández, M., Insuasti Ceballos, H. D., Jiménez Ospina, A., Montenegro Helfer, P., Zapata, E. (2021). Aprovechamiento de datos para la toma de decisiones en el sector público. Caracas: CAF y DNP. Retrieved from https://scioteca.caf.com/handle/123456789/1776

Ezhilarasan, C. & S, R. (2017). Performance prediction using modified clustering techniques with fuzzy association rule mining approach for retail. 2017 International Conference on Intelligent Computing and Control (I2C2). doi: 10.1109/i2c2.2017.8321777

Farhat, J., & Owayjan, M. (2017). ERP Neural Network Inventory Control. Procedia Computer Science, 114, 288-295. https://doi.org/10.1016/j.procs.2017.09.039

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and data mining: towards a unifying framework. In Proceedings of the International Conference on Knowledge Discovery and Data Mining (pp. 82-88).

FENALCO. (2021). La tienda de barrio aliado estratégico de la industria y el consumidor final. Perspectiva de FENALCO frente a la reactivación económica. FENALCO. https://drive.google.com/file/d/1XxP_cD9jQ7BpXw6nuwR_bWieorM-_4bN/view

FENALCO. (2021). La tienda de barrio sigue siendo la joya de la corona para los productos de consumo masivo. FENALCO. https://erp.fenalco.com.co/blog/gremial-4/la-tienda-de-barrio-sigue-siendo-la-joya-de-la-corona-para-los-productos-de-consumo-masivo-456

Fernández-Revuelta Pérez, L., & Romero Blasco, Á. (2022). Un enfoque de ciencia de datos para la toma de decisiones en la estimación de costes - Big Data y aprendizaje automático: A Data Science Approach to Cost Estimation Decision Making - Big Data and Machine Learning. Rev. contab., 25(1), 45-57

Frank, C., Garg, A., Sztandera, L., & Raheja, A. (2003). Forecasting women's apparel sales using mathematical modeling. International Journal of Clothing Science and Technology, 15(2), 107-125. https://doi.org/10.1108/09556220310470097

Franklin Fincowsky, E. B., (2011). Toma de decisiones empresariales. Reseña de "Comportamiento organizacional, enfoque para América Latina" de Franklin, Enrique Benjamín y Krieger, Mario. Contabilidad y Negocios, 6(11), 113-120.

Friedman, J.H., & Meulman, J.J. (2003). Multiple additive regression trees with application in epidemiology. Statistics in Medicine, 22.

Gallagher, C. A., & Watson, H. J. (1988). Métodos cuantitativos para la toma de decisiones en administración (M. González Osuna, Trad.; J. Alonso Cruz, Rev.). McGraw-Hill. (Trabajo original publicado en inglés).

Gallego-Gomez, C., De Pablos-Heredero, C. (2017). Customer relationship management (crm) and big data: a conceptual approach and their impact over the power of data applied to selling strategies. Dyna, 92(3). 274-279. DOI: https://doi.org/10.6036/8071

García Herrero, J., Berlanga de Jesús, A., Molina López, J. M., Patricio Guisado, M. Á., Bustamante, Á. L., & Padilla Arias, W. R. (2018). Ciencia de datos: Técnicas analíticas y aprendizaje estadístico. Altaria.

Gawankar, S. A., Gunasekaran, A., & Kamble, S. (2020). A study on investments in the big data-driven supply chain, performance measures and organisational performance in Indian retail 4.0 context. International Journal of Production Research, 58(5), 1574–1593.

Goyal, D., Goyal, R., Rekha, G., Malik, S., & Tyagi, A. K. (2020). Emerging Trends and Challenges in Data Science and Big Data Analytics. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-8). doi: 10.1109/ic-ETITE47903.2020.316.

Goyal, S., & Modi, N. (2017). Data mining using enhanced decision table classifier for online shopping. 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 313–318. https://doi.org/10.1109/CONFLUENCE.2017.7943168

Gunawan. (2022). Social Commerce from Seller and Region Perspective: A Data Mining for Indonesian E-commerce. In 2022 International Conference on Data Science and Its Applications (ICoDSA) (pp. 268-272). https://doi.org/10.1109/ICoDSA55874.2022.9862835

Gutiérrez Barrera, C. A. (2018). Planes de expansión del comercio minorista: un acercamiento desde la analítica, grandes conjuntos de datos e información de transacciones financieras. Universidad EAFIT.

Han, H., & Trimi, S. (2022). Towards a data science platform for improving SME collaboration through Industry 4.0 technologies. Technological Forecasting and Social Change, 174, 121242. https://doi.org/10.1016/j.techfore.2021.121242

Hanaysha, J.R. (2018). An examination of the factors affecting consumer's purchase decision in the Malaysian retail market, PSU Research Review, Vol. 2 No. 1, pp. 7-23. https://doi.org/10.1108/PRR-08-2017-0034

Harper, M., Mustafina, J., Aljaaf, A. J., Lunn, J., Yasen, S., & Ghali, F. (2019). Data Science Techniques to Support Prediction, Diagnosis and Recode Treatment of Alzheimer's Disease. En 2019 12th International Conference on Developments in eSystems Engineering (DeSE) (pp. 332-339). https://doi.org/10.1109/DeSE.2019.00068

Hassani, H., Beneki, C., Silva, E. S., Vandeput, N., & Madsen, D. Ø. (2021). The science of statistics versus data science: What is the future? Technological Forecasting and Social Change, 173, 121111. https://doi.org/10.1016/j.techfore.2021.121111

Heilman, E., Ganger, R., Coles, J., Ekstrum, J., Hanratty, T., Boslaugh, J., Kendrick, Z., & Ganger, R. (2019). Application of Data Science within the Army Intelligence Warfighting Function: Problem Summary and Key Findings. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 191-195). doi:10.1109/CSCI49370.2019.00039

Henao Rosero, A., & Power, D. J. (2017). State of the Art: Big Data and Business Intelligence. Revista Gestión y Región, (24). ISSN 1900-9771.

Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. Decision Support Systems, 64, 130-141. https://doi.org/10.1016/j.dss.2014.05.013

Hu, X., Yang, Y., Zhu, S., & Chen, L. (2020). Research on a hybrid prediction model for purchase behavior based on logistic regression and support vector machine. En 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 200-204). IEEE. https://doi.org/10.1109/ICAIBD49809.2020.9137484

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. Neurocomputing, 70(1), 489–501.

Hui, H., & Trimi, S. (2022). Towards a data science platform for improving SME collaboration through Industry 4.0 technologies. Technology in Society. https://doi.org/10.1016/j.techfore.2021.121242

Jain, A., Menon, M.N., & Chandra, S. (2015). Sales Forecasting for Retail Chains.

Jain, S., & Kushagra. (2022). Comprehensive Survey on Data science, Lifecycle, Tools and its Research Issues. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (pp. 838-842). doi: 10.1109/COM-IT-CON54601.2022.9850751.

Jansen, R. J. G., Curseu, P. L., Vermeulen, P. A. M., Geurts, J. L. A., & Gibcus, P. (2011). Social capital as a decision aid in strategic decision-making in service organizations. Management Decision, 49(5), 734-747.

Jiménez, R. (2017). Big Data y Machine Learning, los nuevos mejores amigos del marketing. Ctrl: control & strategias, 72-72. http://controlpublicidad.com/big-data-y-machine-learning-nuevos-mejores-amigos-delmarketing/

Ju, C., & Guo, F. (2008). Research and application of customer churn analysis in chain retail industry. In 2008 International Symposium on Electronic Commerce and Security (pp. 670-673). IEEE. https://doi.org/10.1109/ISECS.2008.157

Karimi Dastgerdi, A., & Javdani Gandomani, T. (2021). On the Appropriate Methodologies for Data Science Projects. In 2021 International Conference on Information Technology (ICIT) (pp. 667-673). IEEE. https://doi.org/10.1109/ICIT52682.2021.9491712

Kelleher, J. D., & Tierney, B. (2021). Ciencia de datos: La serie de conocimientos esenciales de MIT Press (ePub). Ediciones UC. https://doi.org/10.9789561427594

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing, 24, 994-1012. https://doi.org/10.1016/j.asoc.2014.08.041

Kopap, A. H., & Elfakharany, E.-E. (2013). Association and Classification Analysis in Retail Case Study. In 2013 23rd International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt (pp. 141-149). https://doi.org/10.1109/ICCTA32607.2013.9529677.

Krstanovic, S., & Paulheim, H. (2017). Ensembles of recurrent neural networks for robust time series forecasting. En M. Bramer y M. Petridis (Eds.), Artificial Intelligence XXXIV (pp. 34–46). Springer International Publishing. ISBN: 978-3-319-71078-5.

Li, C., Hains, M., Wallin, J., & Wu, Q. (2019). Applying data science methods for early prediction of undergraduate student retention. En 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 1337-1340). IEEE. https://doi.org/10.1109/CSCI49370.2019.00250

Liço, L., & Enesi, I. (2021). Performance analysis of neural KNN networks for predicting customer purchases in a real retail department store. Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 1-5. https://doi.org/10.1109/HORA52670.2021.9461316

Linoff, Gordon S., & Berry Michael J.A. (2011). Data mining techniques: for marketing, sales, and customer relationship Management. Indianapolis, IN: Wiley.

Liu, H., Su, B., & Zhang, B. (2007). The Application of Association Rules in Retail Marketing Mix. En 2007 IEEE International Conference on Automation and Logistics (pp. 2514-2517). Jinan, China. DOI: 10.1109/ICAL.2007.4339002.

Lopez, L., & Zuluaga, S. (2013). Impacto de los minimercados en Colombia. Retrieved from https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/76690/1/impacto_minimercados_colombia.pdf.

Lu, J., Yan, Z., Han, J., & Zhang, G. (2019). Data-driven decision-making (D3M): Framework, methodology, and directions. IEEE Transactions on Emerging Topics in Computational Intelligence, 3(4), 286-296. https://doi.org/10.1109/TETCI.2019.2915813

Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. En Advances in Neural Information Processing Systems 30 (pp. 4765-4774).

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. European Journal of Operational Research, 288(1), 111-128. doi: 10.1016/j.ejor.2020.05.038.

Machicao, Jose. (2022). La ciencia de datos para tomar mejores decisiones. 10.13140/RG.2.2.15779.53289.

Marín, G. J., Marcos, P. S., Medina, I. G., & Farias Coelho, P. M. (2020). How Big Data Collected Via Point of Sale Devices in Textile Stores in Spain Resulted in Effective

Online Advertising Targeting. International Journal of Interactive Mobile Technologies (iJIM), 14(13), pp. 65–77. https://doi.org/10.3991/ijim.v14i13.14359

Martínez-Daza, M. A. (2022). The management towards organizational change: The situation of retail shops. VISUAL REVIEW. International Visual Culture Review Revista Internacional De Cultura Visual, 12(4), 1–14. https://doi.org/10.37467/revvisual.v9.3764

Masciari, E., Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise. Journal of Electrical and Computer Engineering, 2019, 8503252. https://doi.org/10.1155/2019/8503252

Mattick, K., Johnston, J., & Croix, A. D. (2018). How to…write a good research question. The Clinical Teacher, 104-108. doi:10.1111/tct.12776

Medina Hernández, E. J. (2021). Analítica: tendencia para optimizar la toma de decisiones a nivel empresarial. Dictamen Libre, (29: Julio-Diciembre). https://doi.org/10.18041/2619-4244/dl.29.7864

Merino Veyl, C. (2015). Modelo de pronóstico de ventas para potenciales locales de una cadena de mejoramiento del hogar. https://repositorio.uchile.cl/handle/2250/132069

Molina Azorín, J. F., López Gamero, M. D., Pereira Moliner, J., Pertusa Ortega, E. M., & Tarí Guilló, J. J. (2012). Métodos híbridos de investigación y dirección de empresas: ventajas e implicaciones. Cuadernos de Economía y Dirección de la Empresa CEDE, 15(2), 55-62. https://doi.org/10.1016/j.cede.2012.01.001

Mompó Serrano, A. (2022). Usos de la Ciencia de datos aplicados al sector Agrícola. Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València

Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. https://doi.org/10.1016/j.crme.2019.11.009

Nithin, S. S. J., Rajasekar, T., Jayanthy, S., Karthik, K., & Rithick, R. R. (2022). Retail demand forecasting using CNN-LSTM model. Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS), 1751–1756. https://doi.org/10.1109/ICEARS53579.2022.9752283

Olson, D.,&Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. International Journal of Forecasting, 19(3), 453–465.

Palkar, A., Deshpande, M., Kalekar, S., & Jaswal, S. (2020). Demand forecasting in retail industry for liquor consumption using LSTM. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 521-525). IEEE. https://doi.org/10.1109/ICESC48915.2020.9155712

Paolotti, D., & Tizzoni, M. (2018). DSAA 2018 Special Session: Data Science for Social Good. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 470-471). IEEE. https://doi.org/10.1109/DSAA.2018.00060

Phyu, M. M., & Khine, M. T. (2023). Retail demand forecasting using sequence to sequence long short-term memory networks. 2023 IEEE Conference on Computer Applications (ICCA), 208-213. https://doi.org/10.1109/ICCA51723.2023.10181450

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data driven decision making. Big Data, 1(1), 51-59. https://doi.org/10.1089/big.2013.1508

Quintero, A., Medina, I., & Rodríguez-Lesmes, P. (2020). Análisis de las dinámicas comerciales de establecimientos de barrio durante el aislamiento obligatorio. https://alianzaefi.com/wp-content/uploads/2023/01/30941.pdf

Rajesh, P. D., Alam, M., Tahernezhadi, M., Vikram, C., & Phaneendra, P. N. (2020). Real Time Data Science Decision Tree Approach to Approve Bank Loan from Lawyer's Perspective. En 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 921-929). doi:10.1109/ICMLA51294.2020.00150

Ramaekers, K., & Janssens, G. K. (2008). On the choice of a demand distribution for inventory management models. European Journal of Industrial Engineering, 2(4), 479–491. http://dx.doi.org/10.1504/EJIE. 2008.018441.

Ranjani, J., Kalaichelvi, V. K. G., Anbalagan, S., S, N. K., & Sudarsan, M. R. (2022). A Deep study of Data science related problems, application and machine learning algorithms utilized in Data science. In 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India (pp. 1-6). https://doi.org/10.1109/IC3IOT53935.2022.9767897.

Razmochaeva, N.V., & Klionskiy, D.M. (2019). Data Presentation and Application of Machine Learning Methods for Automating Retail Sales Management Processes. 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 1444-1448.

Ren, S., Chan, H. L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. Annals of Operations Research, 291(1), 761-777. https://doi.org/10.1007/s10479-019-03148-8

Ren, S., Choi, T.M., & Liu, N. (2015). Fashion sales forecasting with a panel data-based particle-filter model. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(3), 411–421.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. En Proc. ACM SIGKDD International Conference of Knowledge Discovery in Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Ríos, M. E. N., Aldana, E. L., & Madrid, J. R. R. (2004). Evolución del gran comercio minorista en Colombia y sus prácticas contemporáneas adaptativas. Folletos Gerenciales, 8(11).

Riquelme, J.C., Ruiz, R. & Gilbert, K. (2006). Minería de datos: conceptos y tendencias, inteligencia artificial. Revista Iberoamericana de Inteligencia Artificial, 29 pp. 11-18

Rodríguez, J. F. (2016). Implementación de bigdata en las organizaciones como estrategia de aprovechamiento de la información para incorporarla a la cadena de valor del negocio. (Trabajo de grado). Universidad Militar Nueva Granada. http://hdl.handle.net/10654/14411

Rodríguez-Cruz, Y., & Pinto, M. (2018). Modelo de uso de información para la toma de decisiones estratégicas en organizaciones de información. Transinformação, 30(1), 51–64. https://doi.org/10.1590/2318-08892018000100005

Rouhani, S., Ashrafi, A., Zare, A., & Afshari, S. (2016). The impact model of business intelligence on decision support and organizational benefits. Journal of Enterprise-Information Management, 29(1), 19-50. https://doi.org/10.1108/JEIM-12-2014-0126

Ruiz-Lopez, F., Perez-Ortega, J., Ortiz-Hernandez, J., Hernandez-Perez, Y., & Saenz-Sanchez, S. (2021). Systematic review of methodologies in data science. In 2021 Mexican International Conference on Computer Science (ENC) (pp. 1-6). IEEE. https://doi.org/10.1109/ENC53357.2021.9534813

Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K. (2022). Demand Forecasting of a Multinational Retail Company using Deep Learning Frameworks. IFAC-PapersOnLine, 55(10), 395-399. https://doi.org/10.1016/j.ifacol.2022.09.425

Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. (2021). A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance. Information Systems, 99, 101750. https://doi.org/10.1016/j.is.2021.101750

Schnepf, J., Vetter, P., Temel, T., Scheuermann, B., & Schmidt-Thieme, L. (2022). On the Potential of Using ERP Business and System Data for Fraud Detection. En 2022 IEEE International Conference on Big Data (Big Data) (pp. 3081-3091). doi:10.1109/BigData55660.2022.10020785

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2014). Model selection using database characteristics: developing a classification tree for longitudinal incidence data. Marketing Science, 33(2), 188–205. http: //dx.doi.org/10.1287/mksc.2013.0825

Sharma, K., Shetty, A., Jain, A., & Dhanare, R. K. (2021). A Comparative Analysis on Various Business Intelligence (BI), Data Science and Data Analytics Tools. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-11). doi: 10.1109/ICCCI50826.2021.9402226.

Silva Guerra, H. (2012). Panorama del negocio minorista en Colombia. Pensamiento & Gestión, (32), 115-141.

Singh Yadav, N., Goar, V., Singh Yadav, P., Chowdhury, S., Bui, T. N., Thu, N. T., & Vijayakumar, K. (2022). Business Decision making using Data Science. International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1-11). doi:10.1109/ICSES55317.2022.9914352

Singh, A., & Saxena, N. (2021). Data science: Relationship with big data, data driven predictions and machine learning. In 2021 International Conference on Computational Performance Evaluation (ComPE) (pp. 067-072). IEEE. https://doi.org/10.1109/ComPE53109.2021.9752435

Solano-Brenes, A. I. (2013). Toma de decisiones gerenciales. Revista Tecnología En Marcha, 16(3), pág. 44–51 https://revistas.tec.ac.cr/index.php/tec_marcha/article/view/1467

Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 46(1), 411-419. https://doi.org/10.1016/j.dss.2008.07.009.

Taha Falatouri, Farzaneh Darbanian, Patrick Brandtner, & Chibuzor Udokwu. (2022). Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM. Procedia Computer Science, 200, 993-1003. https://doi.org/10.1016/j.procs.2022.01.298

Tandel, T., Wagal, S., Singh, N., Chaudhari, R., & Badgujar, V. (2020). Case Study on an Android App for Inventory Management System with Sales Prediction for Local Shopkeepers in India. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India (pp. 931-934). https://doi.org/10.1109/ICACCS48705.2020.9074234.

Thomassey, S., & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems, 42(1), 408–421. http://dx.doi.org/10.1016/j.dss.2005.01.008.

Thomassey, S., Happiette, M., & Castelain, J. M. (2003). Mean-term textile sales forecasting using families and items classification. Studies in Informatics and Control, 12(1), 41–52.

Toro Ocampo, E. M., et al. (2004). PRONÓSTICO DE VENTAS USANDO REDES NEURONALES. Scientia Et Technica, X(26), 25-30.

Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R. (2022). Classification-based model selection in retail demand forecasting. International Journal of Forecasting, 38(1), 209-223. https://doi.org/10.1016/j.ijforecast.2021.05.010

Van der Voort, H., Bulderen, S. van, Cunningham, S., & Janssen, M. (2021). Data science as knowledge creation: A framework for synergies between data analysts and domain professionals. Technological Forecasting and Social Change, 173, 121160. https://doi.org/10.1016/j.techfore.2021.121160

Van der Voort, H., Klievink, B., Arnaboldi, M., Meijer, A., (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decisionmaking? Gov. Inform. Q. 36 (1), 27–38. https://doi.org/10.1016/j.giq.2018.10.011.

Vandeput, N. (2021). Data science for supply chain forecasting. de Gruyter.

Veres, O., Ilchuk, P., & Kots, O. (2021). Data Science Methods in Project Financing Involvement. In 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT) (pp. 411-414). doi: 10.1109/CSIT52700.2021.9648679.

Wang, Y., Ye, X., & Huo, Y. (2011). Prediction of household food retail prices based on ARIMA Model. In 2011 International Conference on Multimedia Technology (pp. 2301-2305). Hangzhou. doi:10.1109/ICMT.2011.6002376

Wazurkar, P., Bhadoria, R. S., & Bajpai, D. (2017). Predictive analytics in data science for business intelligence solutions. In 2017 7th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 367-370). doi: 10.1109/CSNT.2017.8418568.

Wilson, D. C. et al. (2010). Extreme events, organizations and the politics of strategic decision making. Accounting, Auditing and Accountability Journal, v. 23, n. 5, p. 699-721,

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (Vol. 1, pp. 29-39).

Wu, C., & Xiao, L. (2021). Evidence based on patient's experience data and clinical guidelines-for patient-oriented clinical decision support. In 2021 International Conference on Public Health and Data Science (ICPHDS) (pp. 240-247). doi: 10.1109/ICPHDS53608.2021.00056.

Yelland, P. M., & Dong, X. (2014). Forecasting demand for fashion goods: A hierarchical Bayesian approach. In T. M. Choi, C. L. Hui & Y. Yu (Eds.), Intelligent fashion forecasting systems: Models and applications (pp. 71–94). Springer, Berlin.

Yoo, H., & Pimmel, R. L. (1999). Short term load forecasting using a self-supervised adaptive neural network. IEEE Transactions on Power Systems, 14(2), 779–784.

Yu, H., Cao, L., Li, Y., & Yang, Y. (2011). Research of data mining in electronic commerce. In 2011 International Conference on Consumer Electronics, Communications and Networks (CECNet), Xianning, China (pp. 4323-4326). https://doi.org/10.1109/CECNET.2011.5768320.

Yu, Y., Choi, T.-M., & Hui, C.-L. (2011). An intelligent fast sales forecasting model for fashion products. Expert Systems with Applications, 38 (6), 7373–7379. doi: 10. 1016/j.eswa.2010.12.089

Zampighi, L. M., Kavanau, C. L., & Zampighi, G. A. (2004). The Kohonen self-organizing map: A tool for the clustering and alignment of single particles imaged using random conical tilt. Journal of Structural Biology, 146(3), 368–380.

Zebik, M., Korytkowski, M. , Angryk, R. , & Scherer, R. (2017). Convolutional Neural Networks for Time Series Classification Paper presented at the International Conference on Artificial Intelligence and Soft Computing Cham.

Zheng, Y., Liu, Q. , Chen, E. , Ge, Y. , & Zhao, J. L. (2014). Time Series Classification Us- ing Multi-Channels Deep Convolutional Neural Networks Paper presented at the Web-Age Information Management, Cham.