PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INFRAESTRUCTURA HABITAD Y CREATIVIDAD

MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE



Trabajo De Titulación

Tema:

Modelo predictivo de demanda basado en segmentación de clientes mediante técnicas de clustering y Machine Learning en ventas de productos médicos de laboratorio en el Ecuador

Autor

Bryan Alexis Segura Guamán

DIRECTOR:

Henry Nelson Roa Marin, PhD.

Quito, agosto de 2026

Agradecimientos

Quiero expresar mi profundo agradecimiento a mi familia por el apoyo incondicional, comprensión y motivación durante todo el proceso de formación académica. Su confianza y apoyo emocional fueron fundamentales para afrontar cada desafío y alcanzar esta importante meta. De manera especial, agradezco a mi hija, quien, con su amor y sonrisa, fue mi mayor inspiración y la fuerza que me impulsó a seguir adelante y a culminar este importante logro académico.

Agradezco también al área comercial de la empresa por compartir su experiencia y conocimiento del negocio, que constituyeron un aporte significativo al desarrollo de esta investigación. Sus observaciones y criterios permitieron comprender mejor el contexto empresarial y respaldaron la toma de decisiones en las distintas etapas del proyecto.

Finalmente, expreso mi agradecimiento a mi director de tesis, quien, mediante sus conocimientos, orientación y valiosas observaciones, contribuyó al fortalecimiento de este trabajo de investigación y a mi crecimiento profesional.

ÍNDICE DE CONTENIDO

1. **Introducción** 7
   - 1.1. **Introducción** 7
   - 1.2. **Objetivos** 9
   - 1.3. **Justificación** 9
   - 1.4. **Preguntas de Investigación** 10
2. **Marco Conceptual y Teórico** 11
   - 2.1. **Marco Teórico** 11
     - 2.1.1. **Antecedentes de investigación** 11
     - 2.1.2. **Bases conceptuales** 14
     - 2.1.3. **Teorías, modelos o enfoques** 16
     - 2.1.4. **Fundamentos de la demanda** 19
     - 2.1.5. **Machine Learning aplicado a la predicción** 20
     - 2.1.6. **Tipos de clasificación en machine learning** 21
     - 2.1.7. **Segmentación de clientes** 22
     - 2.1.8. **Técnicas de clustering** 23
     - 2.1.9. **Clustering Jerárquico (Hierarchical Clustering)** 24
     - 2.1.10. **Modelo RFM para análisis de clientes** 26
     - 2.1.11. **Gestión basada en datos** 27
     - 2.1.12. **Métricas de evaluación para modelos de predicción** 28
     - 2.1.13. **Brechas de investigación** 30
     - 2.1.14. **Síntesis del marco teórico** 32
3. **Metodología Aplicada** 34
   - 3.1. **Enfoque y tipo de investigación** 34
   - 3.2. **Marco de trabajo CRISP-ML(Q)** 35
   - 3.3. **Población y Muestra** 37
   - 3.4. **Proceso de Limpieza** 38
   - 3.5. **Preparación de Datos** 39
   - 3.6. **Segmentación de clientes** 39
   - 3.7. **Partición estratificada del dataset** 40
   - 3.8. **Desarrollo de modelos predictivos** 40
   - 3.9. **Estrategia de configuración y ajuste de modelos** 42
   - 3.10. **Regresión Lineal** 42
   - 3.11. **Random Forest** 43

3.12. XGBoost 44
3.13. Evaluación del desempeño de los modelos predictivos 46
4. Análisis y Discusión de Resultados 49
4.1. Análisis exploratorio de datos. 49
4.2. Resultados de la segmentación de clientes. 50
4.3. Resultados de Modelos Predictivos 51
4.3.1. Regresión Lineal 51
4.3.2. Resultados de Random Forest 52
4.3.3. Resultados de XGBoost 53
4.4. Comparación de los modelos predictivos 53
4.5. Discusión de Resultados 55
5. Conclusiones y Recomendaciones 57
5.1. Conclusiones 57
5.2. Recomendaciones 58
Referencias 60

ÍNDICE DE TABLAS

Tabla 1. Variables del Dataset 38
Tabla 2. Distribución clientes por clústers encontrados. 40
Tabla 3. Distribución de la partición del Dataset. 40
Tabla 4. Estadísticos descriptivos de las variables numéricas. 50
Tabla 5. Resultados del modelo de Regresión Lineal. 52
Tabla 6. Resultados de Random Forest. 52
Tabla 7. Resultados de XGBoost. 53
Tabla 8. Comparación del desempeño de los modelos predictivos. 54

ÍNDICE DE ILUSTRACIONES

Ilustración 1. Código Entrenamiento modelo Regresión Lineal. 43
Ilustración 2. Código Entrenamiento Random Forest + Clúster. 44
Ilustración 3. Código Entrenamiento Random Forest sin Clúster. 44
Ilustración 4. Código Entrenamiento XGBoost + Clúster. 45
Ilustración 5. Código Entrenamiento XGBoost sin Clúster. 45
Ilustración 6. MAE (Mean Absolute Error). 46
Ilustración 7. RMSE (Root Mean Squared Error). 47
Ilustración 8. R^2 (Coeficiente de Determinación). 47
Ilustración 9. Método del codo (Elbow Method). 50
Ilustración 10. Silhouette Score. 51

1. Introducción

1.1. Introducción

En los últimos años, el crecimiento acelerado de la información generada por las organizaciones ha impulsado la adopción de la Ciencia de Datos y de técnicas de análisis avanzado como herramientas estratégicas para apoyar la toma de decisiones. La disponibilidad de grandes volúmenes de datos históricos ha permitido que sectores como el comercial, el financiero, el industrial y el de la salud incorporen modelos analíticos capaces de identificar patrones, comprender el comportamiento de los clientes y anticipar escenarios futuros mediante algoritmos de Machine Learning.

En el sector de comercialización de productos médicos de laboratorio, la gestión eficiente de la demanda representa un desafío debido a la diversidad de productos, la variabilidad de los patrones de compra y la necesidad de garantizar la disponibilidad de insumos para la atención de los servicios de salud. Aunque las organizaciones generan información proveniente de sus procesos comerciales, esta suele utilizarse principalmente para análisis descriptivos y reportes históricos, lo que desaprovecha el potencial de las técnicas predictivas para fortalecer la planificación comercial y la gestión basada en datos.

La literatura científica evidencia importantes avances tanto en la clasificación de clientes mediante métodos de agrupamiento como en el desarrollo de modelos predictivos de demanda basados en algoritmos de Machine Learning. Sin embargo, la mayoría de las investigaciones abordan estos procesos de forma independiente, existiendo una limitada evidencia sobre la incorporación de la segmentación de clientes como variable explicativa dentro de modelos predictivos, particularmente en el contexto del sector de productos médicos de laboratorio en Ecuador. Esta situación evidencia una brecha en la investigación que motiva el desarrollo del presente estudio.

En este contexto, la presente investigación tiene como propósito desarrollar un modelo predictivo de demanda basado en la segmentación de clientes mediante técnicas de clustering y algoritmos de Machine Learning aplicados a las ventas de productos médicos de laboratorio en el Ecuador. Para ello, se utilizarán datos históricos de ventas correspondientes al período 2023–2026, a partir de los cuales se identificarán patrones de comportamiento de los clientes mediante las variables RFM (Recency, Frequency y Monetary) y el algoritmo K-Means. Posteriormente, la información obtenida será incorporada como variable explicativa en diferentes modelos predictivos, entre ellos Regresión Lineal, Random Forest y XGBoost, con la finalidad de evaluar y contrastar su rendimiento mediante métricas como MAE y RMSE y el coeficiente de determinación (R^2).

El estudio adopta un enfoque cuantitativo, de alcance correlacional y diseño no experimental, y utiliza técnicas de minería de datos y de aprendizaje automático para analizar la relación entre la segmentación de clientes y la precisión de los modelos predictivos de demanda. Se espera que los resultados contribuyan tanto al ámbito académico como al empresarial, proporcionando evidencia del aporte de integrar técnicas de clustering y Machine Learning en un mismo enfoque analítico para mejorar la toma de decisiones comerciales y la planificación de la demanda.

Finalmente, este documento se estructura en varios capítulos que presentan el problema de investigación, la revisión de la literatura, la metodología empleada, el desarrollo del modelo analítico, el análisis de resultados y las conclusiones del estudio, con el propósito de ofrecer una propuesta metodológica aplicable al contexto de la comercialización de productos médicos de laboratorio en el Ecuador.

1.2. Objetivos

1.2.1 Objetivo general

Desarrollar un modelo predictivo de demanda basado en la segmentación de clientes mediante técnicas de clustering y algoritmos de Machine Learning aplicados a las ventas de productos médicos de laboratorio en el Ecuador.

1.2.2. Objetivos específicos:

1. Identificar patrones de comportamiento de compra mediante técnicas de clustering aplicadas a los clientes.
2. Construir y evaluar modelos predictivos de demanda incorporando la segmentación de clientes como variable explicativa.
3. Comparar el desempeño de los modelos predictivos con y sin segmentación para evaluar su aporte a la precisión de las predicciones.

1.3. Justificación

La presente investigación parte de la necesidad de fortalecer los procesos de toma de decisiones en la gestión de ventas de productos médicos de laboratorio, que actualmente se apoyan en métodos tradicionales que no aprovechan el potencial de los datos disponibles. En un entorno cada vez más competitivo, la capacidad de anticipar la demanda y comprender el comportamiento de los clientes se convierte en un factor clave para la sostenibilidad y el crecimiento de las organizaciones.

Desde la perspectiva del Data Science, este estudio cobra relevancia al integrar técnicas de análisis no supervisado, como el clustering, con modelos predictivos basados en Machine Learning. Esta combinación permite no solo segmentar a los clientes según sus patrones de

compra, sino también incorporar dicha segmentación en la predicción de la demanda, lo que genera modelos más robustos y precisos.

Asimismo, la investigación aporta valor práctico al sector, ya que los resultados obtenidos pueden utilizarse para optimizar las estrategias comerciales, mejorar la planificación de ventas y fortalecer la gestión basada en datos. Esto contribuye a una mayor eficiencia en el uso de recursos y a una mejor adaptación a las dinámicas del mercado.

Finalmente, el desarrollo de este proyecto permite aplicar conocimientos teóricos en un caso real, fortaleciendo competencias en análisis de datos, modelamiento predictivo y toma de decisiones basada en evidencia, aspectos fundamentales en el campo de la ciencia de datos.

A pesar de los avances en la aplicación de técnicas de segmentación de clientes y modelos predictivos en diferentes sectores, existe evidencia limitada sobre la integración de técnicas de clustering y algoritmos de Machine Learning aplicados al sector de productos médicos de laboratorio en el Ecuador. Asimismo, los estudios existentes suelen abordar la segmentación de clientes y la predicción de la demanda de forma independiente, lo que genera una brecha metodológica respecto de la incorporación de la segmentación como variable explicativa en los modelos predictivos de demanda.

1.4. Preguntas de Investigación

¿En qué medida la incorporación de segmentos de clientes obtenidos mediante técnicas de clustering mejora el desempeño de modelos predictivos de demanda construidos a partir de datos históricos de ventas de productos médicos de laboratorio en el Ecuador?

2. Marco Conceptual y Teórico

2.1. Marco Teórico

2.1.1. Antecedentes de investigación

Como punto de partida, se pueden nombrar estudios desde el ámbito internacional, donde Kasem et al. (2024) se centraron en investigar la agrupación de clientes en función del análisis de actualidad, frecuencia y monetario (RFM) y emplearon métricas de validación para derivar agrupaciones óptimas. El algoritmo de agrupamiento K-means, junto con el método del codo, el coeficiente de silueta y el método de estadísticas de brecha, facilita la identificación de distintos segmentos de clientes. El estudio revela tres grupos principales con características únicas: nuevos clientes (grupo A), mejores clientes (grupo B) y clientes intermitentes (grupo C).

En conclusión, la investigación evidenció la importancia de elaborar perfiles de clientes en empresas emergentes digitales mediante el uso de algoritmos de agrupamiento, como K-medias, y de métodos de validación, como Elbow, silueta y Gap Statistics. Los resultados permitieron identificar tres segmentos: nuevos clientes, mejores clientes y clientes intermitentes. Cada grupo requiere estrategias diferenciadas, como soporte personalizado, incentivos exclusivos y promociones dirigidas. Estos hallazgos demuestran que una segmentación adecuada favorece la fidelización, la satisfacción, el crecimiento empresarial y la toma de decisiones alineadas con las necesidades del consumidor.

Por otro lado, Seyedan (2023) se centró en examinar el impacto de la segmentación de clientes en la precisión de la previsión mediante la introducción de un marco novedoso de previsión de la demanda basado en clústeres que aprovecha técnicas de aprendizaje conjunto. Los resultados del estudio demostraron la eficacia del enfoque de conjunto agrupado con errores de previsión mínimos.

Sin embargo, las limitaciones relacionadas con la disponibilidad de datos y la segmentación indican áreas que requieren mayor investigación en futuros estudios.

Los resultados del estudio establecieron que este enfoque no solo reduce los costos de inventario, sino que también mejora la eficiencia de la cadena de suministro. Por lo tanto, el estudio proporcionó información valiosa y metodologías para las empresas que buscan mejorar la precisión de la previsión de la demanda y optimizar las prácticas de gestión de inventario en el sector minorista. Al aprovechar la segmentación de clientes, el aprendizaje profundo conjunto y las técnicas de previsión de la distribución, las organizaciones pueden mejorar los procesos de toma de decisiones, reducir los costos operativos y prosperar en el entorno dinámico de las operaciones de la cadena de suministro.

Desde el contexto latinoamericano, estudios como los de González (2021) han orientado sus esfuerzos al fortalecimiento de los procesos de clasificación, predicción y segmentación de los afiliados de Casur, como parte de sus objetivos estratégicos para mejorar la planificación institucional. Para ello, inicialmente se emplearon modelos basados en reglas y, posteriormente modelos Machine Learning enfocadas en la clasificación y el agrupamiento clustering. Los primeros fueron utilizados como referencia comparativa, mientras que los últimos como base para las etapas posteriores del análisis.

Entre los principales hallazgos del estudio, la regresión logística permitió identificar 9.797 personas clasificadas como potenciales interesados dentro del conjunto general de datos. A partir de este modelo se estimó un costo de adquisición de clientes de $2.608.695, asociado con una probabilidad promedio de venta del 30 % por persona. Por otro lado, mediante el algoritmo K-Means se identificó un segmento con mayor predisposición a adquirir servicios de salud, conformado por 4.727 personas. Este agrupamiento permitió estimar un costo de adquisición de $1.600.000 y una probabilidad media de compra del 45 %.

En esta misma línea, Alonso (2024) desarrolló modelos orientados a la predicción de la demanda como apoyo a la toma de decisiones en la cadena de suministro de una empresa del sector de restaurantes, buscando contribuir a la mejora de su rentabilidad. El estudio consideró variables climáticas, temporales y comerciales relacionadas con el comportamiento de compra. Para la predicción se evaluaron los modelos SARIMAX, Prophet y Random Forest, siendo este último el que alcanzó el mejor desempeño, con un MAPE de 24,9 %. Entre las variables de mayor influencia se identificaron el fin de semana, la temperatura y las festividades. Finalmente, los resultados fueron contrastados con las estrategias de abastecimiento utilizadas por la empresa, evidenciando que estas presentaban una mayor contribución.

El estudio concluyó que, si bien el modelo no ha superado al procedimiento vigente, constituye un avance hacia un proceso de pronóstico de demanda más uniforme, automatizado y basado en datos. Se propone evaluar su aplicación gradual, acompañada de alertas que permitan identificar oportunamente quiebres de stock y pérdidas de inventario. También se recomienda estudiar la influencia de las promociones durante períodos de baja demanda y considerar variables externas que pueden incidir en las ventas, como la temperatura, los feriados y las condiciones meteorológicas. Estas acciones contribuirían a mejorar la planificación y gestión del inventario.

Ahora bien, desde el contexto ecuatoriano, las investigaciones de Pincay (2023), tuvieron como objetivo implementar técnicas avanzadas de minería de datos que faciliten la obtención de información crucial para la toma de decisiones en el Laboratorio Clínico Pura Vida. Situado en Quito, específicamente en el sector La Kennedy, este laboratorio cuenta con una infraestructura tecnológica capaz de realizar 203 pruebas clínicas distintas. La recopilación de los datos financieros y clínicos se realiza mediante dos sistemas informáticos distintos: CLINICAL Lab y SisGem, cada uno especializado en su área.

Se implementaron algoritmos de asociación para descubrir reglas que expliquen el comportamiento del paciente. En particular, se utilizaron los algoritmos A PRIORI y ECLAT, ambos de gran valor para la toma de decisiones empresariales. Este proceso se llevó a cabo mediante programación en Python y se aprovechó BigML para facilitar la identificación ágil de características y patrones en los datos. El análisis resultante no solo proporcionó información relevante, sino que también abrió oportunidades futuras para proyectos de investigación orientados al desarrollo de sistemas informáticos que respalden tanto el diagnóstico como las decisiones de los profesionales sanitarios.

Becerra (2024) desarrolló una investigación orientada a predecir las ventas de una empresa farmacéutica minorista en Ecuador, mediante técnicas de aprendizaje automático y análisis de series temporales. El estudio buscó incrementar la precisión de las proyecciones y proporcionar información útil para fortalecer la toma de decisiones estratégicas dentro de la organización.

El estudio evidenció que el modelo LSTM es una herramienta eficiente para predecir ventas en el sector retail farmacéutico, al captar patrones temporales y estacionales, así como cambios abruptos, y que supera a modelos tradicionales como Random Forest y Holt-Winters. Además, el uso de clustering con K-means permitió segmentar los puntos de venta y personalizar las predicciones, lo que mejoró la gestión de inventarios. Se destaca la importancia de datos limpios y bien procesados para lograr una precisión predictiva. Finalmente, se recomienda implementar un LSTM, capacitar al personal, monitorear el modelo y mantener un entorno flexible.

2.1.2. Bases conceptuales

Demanda: Cantidad de productos que los clientes necesitan o compran en un periodo determinado.

Predicción de la demanda: Estimación de las ventas futuras a partir de datos históricos y patrones de compra.

Modelo predictivo: Herramienta estadística o computacional que anticipa resultados futuros.

Segmentación de clientes: División de los clientes en grupos con características o comportamientos similares.

Clustering: Técnica que agrupa datos o clientes según sus similitudes, sin categorías previas.

Machine Learning: Aprendizaje automático que permite analizar datos y generar predicciones.

Ventas: Proceso de comercialización de productos médicos de laboratorio.

Productos médicos de laboratorio: Insumos, reactivos, equipos y materiales utilizados en el diagnóstico o en análisis clínicos.

Gestión basada en datos: Toma de decisiones mediante el análisis de datos reales.

Estos conceptos son relevantes porque permiten una mejor comprensión del comportamiento comercial de los productos médicos de laboratorio en Ecuador. A partir de los datos históricos de ventas y de la información del mercado, podemos predecir los requisitos futuros para identificar qué productos tendrán alta rotación y cuáles estarán en alta demanda. Esto mejora la planificación empresarial, reduce la incertidumbre en la gestión comercial y hace que la toma de decisiones sobre compras, distribución y servicio al cliente sea más precisa en la industria de la salud.

Este estudio es lógico y beneficioso porque los clientes del sector de la salud no se comportan de manera uniforme, sino que presentan diferentes deseos, frecuencias de compra y capacidades de compra. Necesitan ser evaluados de forma más personalizada para entender sus patrones de consumo de manera distinta. Luego, las empresas pueden adaptar sus

estrategias comerciales según el tipo de cliente, fortalecer la lealtad, centrarse en las áreas más estratégicas y optimizar la entrega oportuna de productos médicos de laboratorio.

Además, el uso de herramientas analíticas en este sector puede generar beneficios prácticos para la gestión empresarial, en particular en el control de inventarios, el suministro y la planificación comercial. Una mala estimación de la demanda en productos médicos de laboratorio puede llevar a pérdidas debido a la caducidad, exceso de stock o falta de disponibilidad. Así, contar con información organizada y modelos de apoyo permite optimizar recursos, brindar una respuesta oportuna al mercado ecuatoriano y fortalecer la competitividad de las empresas del sector de la salud.

2.1.3. Teorías, modelos o enfoques

Como punto de partida, se puede enfatizar la teoría de la demanda; esta teoría representa la cantidad de un bien o servicio que los consumidores están dispuestos y pueden adquirir durante un período determinado. Su comportamiento está influenciado por factores como el precio del producto, el ingreso de los consumidores, los precios de bienes relacionados y sus preferencias. El análisis de la demanda permite comprender el comportamiento de compra y constituye una base para la estimación de ventas futuras (Mankiw, 2021).

También subyace el modelo de predicción de la demanda. Este modelo permite pronosticar el comportamiento futuro de las ventas mediante el análisis de datos históricos, patrones de consumo, ciclos de compra, tendencias y variables externas. En productos de laboratorio médico, su aplicación es importante porque una mala predicción puede provocar desabastecimientos, exceso de inventario, pérdidas por caducidad y retrasos en el servicio. Así, resulta útil para la planificación de inventarios, la optimización de compras, la reducción de costos logísticos y la aseguración oportuna de la disponibilidad de suministros (Sodiq & Joy, 2024).

El análisis predictivo, por su parte, transforma los registros históricos de ventas en información útil para anticipar escenarios futuros y apoyar la toma de decisiones comerciales, logísticas y estratégicas. El análisis predictivo, a diferencia del análisis descriptivo, que explica lo que sucedió, proyecta comportamientos esperados. En el tema propuesto, puede guiarnos en la identificación de productos con mayor demanda potencial, de clientes que podrían aumentar o disminuir sus compras y de patrones comerciales relevantes para mejorar la gestión empresarial basada en datos (Calle et al., 2024).

El aprendizaje automático es un enfoque crítico, ya que permite construir modelos que aprenden patrones a partir de datos sin depender de reglas programadas. Para la predicción de ventas, se pueden utilizar modelos de regresión, Random Forest, XGBoost, Gradient Boosting, KNN, redes neuronales o modelos LSTM. Estos métodos pueden identificar relaciones entre clientes, productos, frecuencia de compra, montos de facturación, estacionalidad y comportamiento histórico, y mejorar la precisión del pronóstico (Saranani et al., 2024).

El enfoque de series temporales ayuda a analizar datos organizados cronológicamente para identificar tendencias, ciclos, estacionalidad y variaciones de demanda a lo largo del tiempo. En las ventas de productos de laboratorio médico, se puede aplicar para estudiar cambios mensuales, trimestrales o estacionales en la compra de reactivos, suministros, equipos o materiales de diagnóstico. Modelos como ARIMA, Holt-Winters, LSTM o GRU complementan el aprendizaje automático al integrar la dimensión temporal del comportamiento comercial (Mariño et al., 2021).

La segmentación de clientes permite dividir el mercado en grupos con características similares, basados en la frecuencia y el volumen de compra, el tipo de producto, el valor monetario, la recurrencia o el comportamiento comercial. Para productos de laboratorio médico, es útil distinguir entre laboratorios clínicos, clínicas privadas, hospitales,

distribuidores y clientes recurrentes, ocasionales o de alto valor. Cuando se combina con modelos predictivos, permite realizar pronósticos y estrategias comerciales más específicos basados en estos segmentos (Ariwiati et al., 2026).

El modelo RFM identifica clientes en función de la recencia, la frecuencia y valor monetario, para encontrar clientes valiosos, frecuentes, nuevos, inactivos o en riesgo. A su vez, el clustering agrupa clientes o productos en función de similitudes internas, junto con algoritmos de agrupamiento como K-means, DBSCAN, clustering jerárquico o modelos de mezcla gaussiana. Ambas técnicas ayudan a mejorar la segmentación y pueden contribuir al diseño de estrategias de lealtad, recuperación, promociones e inventario (Dunaboina, 2024).

Vínculos con clientes y la minería de datos constituyen el CRM que permite gestionar las relaciones con los clientes mediante información sobre el historial de compras, preferencias, frecuencia de interacción y valor generado. El CRM puede utilizarse en una empresa dedicada a productos de laboratorio médico para generar modelos de segmentación y predicción. Por otro lado, la minería de datos permite descubrir patrones ocultos en bases de datos de facturación, inventario y ventas pasadas e identificar productos de alta rotación, tendencias, oportunidades de venta cruzada y comportamientos comerciales similares.

También están la cadena de suministro inteligente y la toma de decisiones basada en datos. Esta aplicación combina inteligencia artificial, predicción de la demanda y análisis de datos para optimizar las compras, los inventarios, la distribución y la reposición de productos. Es importante en el sector médico para evitar la escasez de stock y el exceso de inventario.

Finalmente, la toma de decisiones basada en datos resume todos los enfoques anteriores; ayuda a definir qué productos adquirir, cuánto inventario mantener, a qué clientes dirigirse y qué cambiará en la demanda del mercado ecuatoriano (Dias, 2025).

2.1.4. Fundamentos de la demanda

La demanda se refiere a la cantidad de bienes y/o servicios que los consumidores están dispuestos a adquirir para satisfacer sus necesidades o deseos. Estos compradores no solo tienen el deseo, sino también la capacidad económica para llevar a cabo la transacción a un precio específico y en un lugar determinado. Además, representa el volumen de productos o servicios que una persona está dispuesta a adquirir bajo ciertas condiciones de precio y ubicación, lo que le permite cubrir total o parcialmente sus requerimientos específicos o disfrutar del valor intrínseco que estos ofrecen (Syamril et al., 2024).

La demanda de un producto puede variar en función de factores como su precio, calidad percibida, publicidad, nivel de ingresos y confianza del consumidor, así como las tendencias de moda y los cambios en los gustos. Analizar los diversos factores y la complejidad del mercado de bienes de consumo envasados que inciden en la demanda resulta de gran utilidad (Wu, 2023).

La demanda de un producto o servicio está influida por factores económicos y sociales. Los principales son el precio, el ingreso del consumidor, las preferencias, los bienes relacionados, las expectativas futuras y el tamaño de la población. Si un producto es de buena calidad y rentable, la demanda aumenta; si es caro o de baja calidad, disminuye. Además, el ingreso afecta el poder adquisitivo y los gustos y hábitos llevan a la preferencia por ciertos bienes o servicios (Hanaysha et al., 2021).

La oferta representa el volumen de productos o servicios que los productores pueden y desean colocar en el mercado a distintos precios dentro de un período específico. Su comportamiento está condicionado por elementos como los costos de producción, la tecnología, el valor de los insumos, las expectativas empresariales y las regulaciones gubernamentales. Manteniendo constantes los demás factores, un incremento en los precios

suele generar mayores incentivos para que los productores aumenten la cantidad de bienes disponibles en el mercado (Mankiw, 2021).

La oferta y la demanda son aspectos fundamentales del comportamiento del mercado. Los dos están inextricablemente vinculados porque el precio, el ingreso, la producción, la tecnología o las políticas públicas pueden modificar las decisiones de los consumidores o de los productores. Por lo tanto, analizar estos aspectos ayudará a las empresas, los gobiernos y las personas a tomar mejores decisiones económicas. En el contexto de los productos médicos de laboratorio, comprender estos factores es esencial para anticipar la demanda, gestionar inventarios y diseñar estrategias comerciales más eficientes (Syamril et al., 2024).

2.1.5. Machine Learning aplicado a la predicción

El machine learning es una rama de la inteligencia artificial que realiza análisis de datos sin necesidad de instrucciones directas. Esta tecnología es capaz de manejar volúmenes significativos de información histórica, reconociendo patrones y anticipando conexiones entre datos antes no identificados. Además, puede ejecutar tareas tanto de clasificación como predicción en diversos formatos, incluyendo documentos, imágenes y números (Pertuz, 2022).

Los sistemas de predicción que emplean algoritmos de Machine Learning, o aprendizaje automático, permiten a las empresas generar estimaciones y pronósticos a partir de los datos provenientes del comportamiento de sus usuarios. Esto no solo optimiza procesos, sino que también genera un valor significativo para las organizaciones. El aprendizaje automático, como uno de los componentes fundamentales del ámbito más amplio de la inteligencia artificial (IA), se enfoca en desarrollar sistemas que pueden aprender y mejorar su rendimiento a partir de la información que se les proporciona (Forero & Negre, 2023).

Es crucial entender que el término "aprendizaje automático" suele ir acompañado de "Inteligencia Artificial". Sin embargo, aunque ambos conceptos están íntimamente conectados,

no son sinónimos. En este sentido, podemos afirmar que todos los sistemas diseñados bajo el concepto de Machine Learning forman parte del desarrollo generalizado conocido como inteligencia artificial; pero también es válido señalar que no todas las aplicaciones clasificadas como inteligencia artificial corresponden necesariamente a modelos específicos de Machine Learning (Hermitaño, 2022).

2.1.6. Tipos de clasificación en machine learning

En aprendizaje automático, existen cuatro tipos principales de clasificación: binaria, multiclase, multietiqueta y con clases desbalanceadas, las cuales, según León (2025), se desarrollan de la siguiente manera:

- **Clasificación Binaria:** Se clasifica en dos categorías mutuamente excluyentes, utilizando un conjunto de datos etiquetado como verdadero/falso u otras combinaciones. Algoritmos como Logistic Regression y Support Vector Machines (SVM) son específicos para esta tarea, aunque K-Nearest Neighbors (K-NN) y Árboles de Decisión también pueden aplicarse.
- **Clasificación Multiclase:** Involucra al menos tres clases mutuamente excluyentes y busca predecir a qué clase pertenece un ejemplo. Algoritmos como Random Forest, Naive Bayes, K-NN, y otros pueden adaptarse a este enfoque mediante estrategias como:
    - **One-vs-One (OvO)**: Se entrena un clasificador para cada par de clases.
    - **One-vs-Rest (OvR)**: Se entrena un clasificador por cada clase, considerando las demás como clases negativas.
- **Clasificación Multietiqueta**: Predice 0 o más etiquetas por entrada, lo que permite que un ejemplo esté asociado a múltiples etiquetas. Se usa en tareas como el etiquetado en el

procesamiento del lenguaje natural. Modelos especializados incluyen Multi-label Decision Trees y Multi-label Random Forests.

- **Clasificación con Clases Desbalanceadas**: Aquí, las clases tienen distribuciones desiguales de ejemplos, lo que dificulta el aprendizaje de las clases minoritarias. Problemas típicos incluyen la detección de fraudes y el diagnóstico de enfermedades raras. Los modelos convencionales tienden a favorecer a la clase mayoritaria, ignorando a las minorías.

2.1.7. Segmentación de clientes

La segmentación de clientes consiste en clasificar a los consumidores según rasgos, necesidades y actitudes compartidas. Este enfoque permite a las marcas concentrar sus esfuerzos estratégicamente en aquellos grupos que resultan más rentables. Identificar estos segmentos es crucial para crear campañas publicitarias efectivas, captar clientes con alto potencial de rentabilidad y fomentar la lealtad del consumidor. Normalmente, esta clasificación se realiza basándose en características comunes, como la personalidad o las preferencias, que están directamente vinculadas con comportamientos que generan ganancias significativas para el negocio (López, 2022).

La segmentación de clientes permite a las empresas enfocarse eficazmente en su público objetivo y personalizar estrategias enfocadas en responder a las demandas particulares de cada cliente. Este enfoque es fundamental en el marketing centrado en el consumidor. A continuación, se presentan ocho modelos de segmentación que ofrecen una comprensión más profunda sobre el comportamiento y preferencias de los consumidores, lo que no solo optimiza la experiencia del cliente, sino que también incrementa significativamente el valor durante todo su ciclo de vida (Muñoz et al., 2022).

La clasificación de los clientes constituye un elemento clave para diseñar estrategias de marketing actuales y eficientes. Al comprender los patrones de comportamiento del consumidor, las organizaciones pueden enfocar sus iniciativas promocionales con mayor precisión. Este enfoque permite a las empresas adaptar sus tácticas comerciales sin necesidad de elaborar un plan específico para cada cliente. Como resultado, no solo se enriquece la experiencia general del usuario, sino que también se potencia el retorno sobre la inversión (ROI) total de la empresa (Rojo, 2024).

2.1.8. Técnicas de clustering

Los algoritmos de agrupamiento se han consolidado como herramientas relevantes para detectar patrones en diferentes contextos espaciales, facilitando el análisis de grandes cantidades de información y extracción crucial que apoya la toma efectiva de decisiones. El proceso de clustering, que simula cómo los humanos interpretan visualmente los datos, desempeña un papel vital en el aprendizaje automático y se aplica a numerosas disciplinas que manejan información. Su relevancia se refleja claramente en campos tan diversos como la ciberseguridad, el procesamiento documental, la detección de fraudes y la biología molecular, lo que subraya su importancia para examinar y entender patrones significativos (Mohammad, 2023).

Los algoritmos de clustering han demostrado ser herramientas valiosas para optimizar la cadena de suministro, ya que permiten agrupar ubicaciones o proveedores con características similares, lo que se traduce en una mejora notable de la eficiencia logística. Dentro del área de investigación y desarrollo de productos, el uso del clustering facilita la identificación de preferencias y tendencias entre los consumidores, dirigiendo el diseño hacia soluciones acordes con las demandas del mercado (Becerra, 2024).

Este enfoque tecnológico representa una evolución relevante de los sistemas actuales de información geográfica, al incorporar herramientas avanzadas que fortalecen la administración y el procesamiento de información espacial. La esencia del clustering radica en evaluar la similitud entre observaciones para formar grupos; según el tipo de datos a analizar, se implementan diferentes técnicas adaptadas a cada situación (Castro et al., 2025).

2.1.9. Clustering Jerárquico (Hierarchical Clustering)

El clustering jerárquico es un método esencial para la agrupación de datos, basado en las distancias entre ellos. Su objetivo principal es asegurar que los elementos de cada clúster sean lo más homogéneos posible. Este enfoque se puede dividir en dos variantes: aglomerativa, donde cada punto inicial se considera un clúster independiente y luego se combinan progresivamente; y divisiva, que comienza con un único clúster y lo fragmenta gradualmente en clústeres más pequeños (Vichi et al., 2022).

Los métodos jerárquicos son altamente valorados por su capacidad para construir una estructura jerárquica a partir del conjunto de datos estudiado. El proceso aglomerativo consiste en fusionar pares consecutivos de objetos o clústeres hasta obtener un único clúster. En contraste, el método divisivo descompone los grupos existentes hasta que cada objeto queda contenido en su propio clúster (Vichi et al., 2022).

a) Mean Shift

El algoritmo *Mean Shift* constituye una técnica iterativa diseñada para identificar áreas densamente pobladas de puntos de datos. Esta técnica utiliza el concepto de Kernel Density Estimation (KDE), asumiendo que dichos puntos provienen de una distribución desconocida, lo que permite estimar la distribución subyacente.

Este algoritmo inicia colocando una ventana circular móvil centrada en un punto aleatorio e itera desplazándola hacia zonas con mayor concentración de datos. Este movimiento continuará hasta alcanzar la convergencia cuando ya no haya cambios significativos en la cantidad total encontrada dentro del área considerada y, finalmente, agrupará los puntos según su ubicación relativa a esta ventana deslizante (Castro et al., 2025).

b) K-Means

Por otro lado, el algoritmo *K-Means* opera dividiendo conjuntos específicos en "K" grupos predeterminados. El objetivo consiste en reducir al mínimo la distancia acumulada entre cada observación y el centroide correspondiente, es decir, el punto central de cada grupo identificado por el propio algoritmo, que asigna cada punto de datos al centroide más cercano y recalcula continuamente estos centroides hasta que se estabilizan, sin variaciones adicionales.

Esto empieza determinando cuántos clústeres deben formarse inicialmente, seleccionando k puntos al azar como centros iniciales, antes de asignar, conforme corresponda, cada dato individualmente a sus respectivos clústeres; tras esto, procede a recalcular nuevamente los centroides, repitiendo todo el ciclo continuamente hasta llegar al estado deseado, conocido como convergencia (Marrero et al., 2021).

c) DBSCAN

Con respecto al algoritmo *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*), este enfoque distintivo identifica clústers basándose exclusivamente en cómo están distribuidos esos mismos datos sin requerir definir previamente cuantos serán estos grupos; trabaja utilizando dos parámetros cruciales: épsilon (eps), marcando distancia máxima cualitativa para juzgar si dos elementos pertenecen juntos bajo el mismo grupo u otro distinto mediante criterios establecidos además del número mínimo requerido cercano (min_samples) necesario para considerar válida cualquier formación grupal.

DBSCAN identifica segmentos densamente concentrados, mientras trata de forma aislada otros menos densos, catalogándolos como ruido o outliers. Es particularmente efectivo al buscar patrones arbitrarios presentes incluso en grandes volúmenes de datos complejos (Singh et al., 2022).

2.1.10. Modelo RFM para análisis de clientes

La segmentación RFM es un enfoque sencillo, pero altamente eficaz que se utiliza en diversas áreas del mercado. Este método de análisis aprovecha la información sobre el comportamiento de compra de los consumidores durante un período anterior, fundamentándose en tres variables clave que originan su nombre: (R) Recencia, que indica el tiempo transcurrido desde la última compra realizada por el consumidor hasta hoy; (F) Frecuencia, que mide cuántas transacciones ha efectuado un cliente dentro de un marco temporal específico; y (M) Monto de compra, correspondiente al total gastado por el cliente durante ese mismo periodo (Mellet et al., 2026).

El modelo RFM constituye una metodología práctica y eficaz para segmentar clientes. Esto se debe a su base en variables transaccionales simples, que poseen un notable poder predictivo. Esta herramienta es accesible para cualquier tipo de organización, ya que permite clasificar a los clientes según su probabilidad de recompra y su aporte económico.

Una ventaja del modelo RFM radica en que no demanda inversiones significativas en infraestructura tecnológica para proporcionar resultados claros y accionables, fácilmente comprensibles para los equipos comerciales. Además, los consumidores con compras recientes y frecuentes, incluso si son de compras moderadas, presentan un mayor potencial de fidelización frente a aquellos que realizan transacciones de mayor valor, pero con menor frecuencia. Estos hallazgos resultan cruciales, especialmente para sectores como el ciclismo o el fitness, donde la sostenibilidad económica depende en gran medida de la recurrencia en

servicios y accesorios. Al adoptar el enfoque RFM, estas industrias pueden optimizar sus estrategias comerciales para lograr una rentabilidad sostenida (Diez, 2022).

2.1.11. Gestión basada en datos

La gestión basada en datos representa un enfoque organizacional que fundamenta la toma de decisiones en el análisis sistemático de información, sustituyendo la intuición por evidencia empírica. Su adopción permite a las organizaciones mejorar la precisión de las decisiones, optimizar procesos, identificar oportunidades de mejora y responder de manera más eficiente a los cambios del entorno competitivo. Además, la disponibilidad de herramientas analíticas facilita la generación de información valiosa para apoyar la formulación de estrategias y el seguimiento del desempeño empresarial (Prakash, 2024).

Primordialmente, la gestión basada en datos sienta cimientos mucho más sólidos para el proceso decisional; al basarse en datos precisos, los líderes pueden tener mayor confianza en que sus decisiones conducirán a resultados favorables. Además, la agilidad que ofrece este tipo de toma de decisiones es inigualable: permite procesar y analizar la información con rapidez. Este aspecto es especialmente crucial en el dinámico entorno empresarial actual, donde cada segunda cuenta y puede marcar la diferencia entre el éxito y el fracaso. Adoptar una estrategia basada en datos no solo optimiza las decisiones empresariales, sino que también posiciona a las organizaciones para prosperar ante los desafíos constantes e impredecibles del mercado contemporáneo (Prakash, 2024).

En este sentido, cobra especial relevancia la toma de decisiones basada en datos (data-driven decision making), entendida como un enfoque gerencial que utiliza información obtenida de diversas fuentes para respaldar la planificación, la evaluación de alternativas y la definición de estrategias organizacionales. A través del análisis sistemático de datos históricos y actuales, las empresas pueden identificar patrones, anticipar tendencias, optimizar procesos

y reducir la incertidumbre asociada a la toma de decisiones. Este enfoque favorece una gestión más objetiva y alineada con los objetivos estratégicos de la organización, contribuyendo a mejorar el desempeño, la eficiencia operativa y la capacidad de respuesta ante cambios del entorno empresarial (Abdul-Azeez et al., 2024).

2.1.12. Métricas de evaluación para modelos de predicción

Las métricas de evaluación son métodos estadísticos que permiten medir el rendimiento de un modelo predictivo. En el caso de un modelo de predicción de demanda, estas métricas permiten comparar las ventas reales con los valores predichos por el modelo. Es muy importante utilizarlas en la toma de decisiones empresariales, porque no basta con construir un modelo de aprendizaje automático; también es necesario verificar que los resultados sean fiables, útiles y adecuados para la toma de decisiones empresariales (Nasseri et al., 2023).

En las ventas de productos médicos de laboratorio, las métricas de evaluación ayudan a determinar si el modelo es acertado respecto de la demanda de un producto, cliente o segmento en particular. Esto ayuda a minimizar los errores en la planificación de inventario, a evitar desabastecimientos, a disminuir el exceso de stock y a mejorar la gestión comercial. Las métricas más comunes utilizadas en modelos de regresión y predicción de demanda son el Error Absoluto Medio, el Error Cuadrático Medio, la Raíz del Error Cuadrático Medio y el Coeficiente de Determinación (Aldahmani et al., 2024).

a) Error Absoluto Medio

El error absoluto medio (MAE) representa una métrica que permite determinar el promedio de las diferencias, en términos absolutos, entre los valores observados y las estimaciones generadas por el modelo. Esta medida indica, en general, cuánto se equivoca el modelo en sus predicciones. Al usar valores absolutos, los errores positivos y negativos se cancelan entre sí, lo que permite observar el error promedio.

En esta fórmula y_i es el valor real, $\hat{y}_i$ el valor predicho y n el número total de observaciones. Un MAE bajo en un modelo predictivo de demanda indica que las predicciones se ajustan bien a las ventas reales. Por ejemplo, si el modelo predice la demanda mensual de reactivos, suministros o equipos médicos, el MAE permite conocer el error promedio en unidades vendidas o en valores monetarios. Su principal ventaja es que es facilidad de interpretación debido a que mantiene la unidad de medida de la variable estudiada.

b) Error Cuadrático Medio

El error cuadrático medio (MSE) es el promedio de los cuadrados de los errores entre los valores reales y los estimados. Al igual que el MAE, esta métrica penaliza más severamente los errores grandes, ya que cada diferencia se eleva al cuadrado. Por lo tanto, resulta útil cuando se quiere saber si el modelo comete errores grandes en algunas predicciones (Nasseri et al., 2023).

En términos de ventas de productos médicos de laboratorio, el MSE permite evaluar si el modelo subestima la demanda de ciertos productos o de determinados clientes. Si un producto tiene alta rotación y el modelo subestima la demanda, el error cuadrático será elevado. Esta métrica puede ser útil para comparar diferentes algoritmos de aprendizaje automático, como Random Forest, regresión lineal, árboles de decisión y redes neuronales. El resultado en unidades al cuadrado es menos directo de interpretar (Aldahmani et al., 2024).

c) Error Cuadrático Medio de la Raíz

El error cuadrático medio de la raíz (RMSE) se define como la raíz cuadrada del error cuadrático medio. Esta métrica tiene la misma capacidad de penalizar errores grandes que el MSE, mantiene la misma unidad de medida que la variable dependiente analizada.. Por lo tanto, es una de las métricas más populares en problemas de predicción de la demanda y de las ventas (Ranabhatt et al., 2025).

En un modelo de demanda para productos médicos de laboratorio, el RMSE nos permite saber cuánto, en promedio, la predicción se desvía de la demanda real, asignando más peso a los errores más grandes. Si el RMSE es bajo, significa que el modelo tiene una buena capacidad predictiva. Si es alto, significa que el modelo no está captando patrones de consumo, estacionalidad, comportamiento del cliente ni variaciones en las ventas. Esta métrica es particularmente valiosa cuando necesitamos comparar la precisión de los distintos modelos y elegir el más adecuado (Nivela, 2024).

d) Coeficiente de Determinación

El coeficiente de determinación R^2 indica qué proporción de la variación observada en la variable dependiente puede ser explicada por el modelo empleado. En otras palabras, nos permite evaluar qué tan bien el modelo describe el comportamiento de la demanda en función de las variables utilizadas, como el historial de ventas, el tipo de cliente, la frecuencia de compra, el segmento, el precio, los productos adquiridos y la ubicación geográfica (Nasseri et al., 2023).

El valor de R^2 suele estar entre 0 y 1. Un valor cercano a 1 indica que el modelo explica la mayor parte de la variación de la demanda, mientras que un valor cercano a 0 indica que no la explica. En el caso de productos médicos de laboratorio, un R^2 alto indica que las variables seleccionadas nos permitirán comprender el comportamiento de compra del cliente. Sin embargo, esta métrica no debe analizarse de forma aislada porque un modelo puede tener un buen R^2 pero en ciertos casos hay errores muy altos. Por lo tanto, debemos ser capaces de interpretarlo junto con MAE, MSE y RMSE (Aldahmani et al., 2024).

2.1.13. Brechas de investigación

Como punto de partida, la mayoría de la literatura se centra en el comercio minorista general, el comercio electrónico, los productos farmacéuticos y las cadenas de suministro

hospitalarias, pero no aborda realmente las dinámicas comerciales de los suministros de laboratorio ni por qué la demanda puede depender de la estacionalidad epidemiológica, la rotación institucional, las compras recurrentes, las emergencias clínicas, la disponibilidad de inventario y el comportamiento del cliente. Aunque el análisis predictivo mejora la gestión de inventarios y previene la falta de existencias en el sector salud, hay poca evidencia empírica sobre su aplicación en las empresas de distribución médica ecuatorianas (Nasseri et al., 2023).

Otra brecha es que la segmentación de clientes y los modelos predictivos de demanda no están integrados adecuadamente. Algunos estudios han realizado agrupamientos, RFM o machine learning para la segmentación de clientes, mientras que otros han aplicado modelos de predicción de ventas por separado. Pero son pocos los trabajos que han combinado ambos pasos en un solo modelo: primero segmentando a los clientes según su comportamiento de compra y luego entrenando un modelo de predicción por segmento. Esto es importante porque los clientes de productos médicos no presentan el mismo patrón en cuanto a volúmenes, frecuencias y ciclos de demanda. Un estudio reciente muestra que los modelos híbridos con agrupamiento tienen una mayor precisión que los modelos generales, por lo que deberíamos aplicar esta lógica a la industria médica (Thiyagarajan et al., 2026).

De igual manera, aunque los modelos de predicción son altamente precisos, solo lo son si los datos están limpios, completos, normalizados y representativos. En las empresas de productos médicos de laboratorio, los datos pueden ser inconsistentes debido a códigos de producto erróneos, compras irregulares, clientes inactivos, datos faltantes, devoluciones, descuentos, faltas de stock o variaciones estacionales en los códigos de producto. Además, numerosos estudios recientes han advertido de que los modelos tradicionales podrían no ser capaces de gestionar demandas intermitentes, heterogéneas o influidas por factores externos. Por lo tanto, es importante desarrollar modelos con características predictivas que no solo se apliquen a las ventas, sino que también puedan usarse para predecir datos, analizar clústeres y

comparar algoritmos, y evaluar los resultados con métricas apropiadas para mejorar la predicción del comercio y el inventario (Eglite & Birzniece, 2022).

2.1.14. Síntesis del marco teórico

En función de este compendio teórico desarrollado, se puede enfatizar que estos argumentos investigativos, pueden ayudar a establecer que la predicción de la demanda, la segmentación de clientes, el clustering y el aprendizaje automático son métodos útiles para mejorar la gestión comercial en sectores donde el comportamiento de compra es variable y depende de múltiples factores. En el ámbito de la literatura, estas estrategias se han utilizado en contextos internacionales, latinoamericanos y ecuatorianos para optimizar las ventas, los inventarios, las cadenas de suministro y las decisiones empresariales. En este sentido, la literatura evidencia cómo podemos utilizar modelos basados en datos históricos, patrones de consumo y características de los clientes para prever mejor la demanda futura.

De manera similar, los fundamentos teóricos muestran que la demanda de productos médicos de laboratorio puede verse influida por la frecuencia de compra, el tipo de cliente, la rotación de productos, la estacionalidad, los precios, la disponibilidad, las condiciones sanitarias y las necesidades institucionales. Por lo tanto, el método propuesto se justifica por la integración de técnicas de segmentación y de aprendizaje automático, ya que no todos los clientes exhiben el mismo comportamiento comercial. El clustering ayudará a agrupar a los clientes con características similares, mientras que los modelos predictivos ayudarán a predecir la demanda futura a partir de los patrones de cada segmento.

El marco teórico también sugiere la relevancia de desarrollar un modelo predictivo de demanda para las ventas de productos médicos de laboratorio en Ecuador, a fin de abordar una brecha en la literatura: la falta de segmentación de clientes y de predicción de la demanda en este sector específico, que es el único con evidencia que respalda su aplicación. La

investigación debería proporcionar a la empresa una herramienta analítica para apoyar la planificación de inventarios y la gestión de existencias, evitar el exceso de stock, minimizar las pérdidas por caducidad y mejorar la toma de decisiones comerciales. Así, el estudio no solo tiene una base cuantitativa, sino que también puede ser útil para organizaciones del sector salud.

3. Metodología Aplicada

3.1. Enfoque y tipo de investigación

La presente investigación se fundamenta en un enfoque cuantitativo**,** enmarcado en el paradigma empírico-analítico, ya que utiliza técnicas estadísticas y algoritmos de Machine Learning para analizar datos históricos de ventas y desarrollar un modelo predictivo de demanda. Este enfoque permite medir objetivamente el comportamiento de las variables, identificar patrones de compra y evaluar el desempeño de los modelos mediante métricas cuantitativas, garantizando la reproducibilidad del proceso de investigación (Mitchell, 1997).

En cuanto a su finalidad, el estudio es de tipo aplicado, debido a que busca desarrollar una solución analítica que contribuya a mejorar la predicción de la demanda de productos médicos de laboratorio y sirva como herramienta de apoyo para la toma de decisiones comerciales. Para ello, se emplean registros históricos reales de ventas correspondientes a una empresa del sector, generando conocimiento con aplicación práctica en un entorno organizacional.

Desde la perspectiva del diseño, la investigación es no experimental, ya que las variables son observadas en su contexto natural sin manipulación deliberada por parte del investigador. Asimismo, presenta un diseño retrospectivo, al utilizar información histórica comprendida entre enero de 2023 y marzo de 2026 para analizar el comportamiento de la demanda y construir los modelos predictivos.

Respecto a su alcance, la investigación es de tipo descriptivo-predictivo. Es descriptiva debido a que analiza el comportamiento de compra de los clientes mediante aplicación de técnicas de segmentación, e incorpora un enfoque predictivo al desarrollar y evaluar modelos de Machine Learning para estimar la demanda de productos médicos de laboratorio.

Finalmente, el desarrollo de la investigación se estructura bajo la metodología CRISP-ML(Q)**,** seleccionada por proporcionar un marco de trabajo sistemático para el desarrollo de proyectos de Machine Learning, integrando actividades de comprensión del problema, preparación y aseguramiento de la calidad de los datos, modelado, evaluación y validación de los resultados, garantizando la trazabilidad y calidad del proceso analítico (Studer et al., 2021).

3.2. Marco de trabajo CRISP-ML(Q)

El diseño metodológico de la presente investigación se estructura en el marco de referencia CRISP-ML(Q) (Cross-Industry Standard Process for Machine Learning with Quality Assurance) es una metodología para el desarrollo de proyectos de Machine Learning que incorpora principios de aseguramiento de la calidad durante todo el ciclo de vida del modelo, desde la comprensión del problema hasta su evaluación e implementación (Studer et al., 2021).

Este diseño permite organizar el proceso de investigación en etapas claramente definidas, facilitando el desarrollo sistemático de proyectos de Machine Learning e incorporando criterios de aseguramiento de la calidad durante todo el proceso. El proceso metodológico adoptado en la presente investigación se desarrolla a través de las siguientes fases:

a) **Entendimiento del problema:** En esta etapa inicial se definió el problema de investigación como el desarrollo de un modelo predictivo de demanda para las ventas de productos médicos de laboratorio, utilizando datos históricos correspondientes al período 2023–2026. Asimismo, se estableció como objetivo evaluar la influencia de la segmentación de clientes, obtenida mediante técnicas de clustering, en la precisión de los modelos de Machine Learning empleados para la predicción de la demanda (Studer et al., 2021).

b) **Entendimiento y análisis de datos:** En esta fase se realizó la exploración inicial de la base de datos con el propósito de comprender su estructura, identificar las variables relevantes y analizar la calidad de la información disponible. Además, se efectuó un análisis exploratorio de los datos para identificar patrones de comportamiento, distribución de las variables, posibles valores atípicos y relaciones entre clientes, productos y ventas, proporcionando una base sólida para las etapas posteriores del proceso analítico (Han et al., 2012).

c) **Preparación de los datos:** Durante esta etapa se llevaron a cabo procesos de limpieza, depuración, transformación e integración de los datos históricos de ventas. Posteriormente, se construyeron las variables RFM (Recency, Frequency y Monetary) para caracterizar el comportamiento de compra de los clientes y se aplicaron técnicas de escalamiento de datos como paso previo al proceso de segmentación mediante el algoritmo K-Means. Asimismo, se preparó el conjunto de datos que serviría como entrada para los modelos predictivos de demanda (Studer et al., 2021).

d) **Modelado:** En esta fase se desarrolló el proceso de segmentación de clientes utilizando el algoritmo K-Means con base en las variables RFM. Una vez obtenidos los segmentos, estos fueron incorporados como variable explicativa dentro de los modelos predictivos de demanda. Posteriormente, se implementaron y entrenaron diferentes algoritmos de Machine Learning supervisado, entre ellos Regresión Lineal, Random Forest y XGBoost, con el objetivo de comparar su capacidad predictiva e identificar el modelo con mejor desempeño (James et al., 2021).

e) **Evaluación:** En esta etapa se evaluó el desempeño de los modelos predictivos mediante métricas de regresión como el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el coeficiente de determinación (R^2). Adicionalmente, se compararon los resultados obtenidos por los modelos construidos con y sin la incorporación de la

segmentación de clientes (clúster), con el propósito de determinar el aporte de esta variable al desempeño predictivo (Géron, 2022).

f) **Integración Estratégica:** En esta se analizará la utilidad práctica de los resultados obtenidos para apoyar la toma de decisiones en el área comercial. La información generada permitirá identificar segmentos de clientes con comportamientos diferenciados y mejorar la estimación de la demanda de productos médicos de laboratorio, proporcionando una herramienta analítica que contribuya a optimizar la planificación comercial y fortalecer la toma de decisiones basada en datos (Studer et al., 2021).

3.3. Población y Muestra

La población de estudio está compuesto por los registros históricos de ventas de productos médicos de laboratorio correspondientes a una empresa del sector en el Ecuador. El universo de análisis comprende un total de 67.326 registros recopilados entre enero de 2023 y marzo de 2026, los cuales contienen información relacionada con clientes, productos y transacciones comerciales. Estas variables permiten caracterizar el comportamiento de consumo de los clientes, analizar la demanda de productos y construir los modelos de segmentación y predicción propuestos en la investigación.

La información utilizada constituye la totalidad de los registros disponibles para el período de estudio, por lo que no fue necesario aplicar técnicas de muestreo. El conjunto de datos incluye 21 variables asociadas a clientes, productos y transacciones comerciales, las cuales se describen en la siguiente tabla:

Tabla 1. Variables del Dataset

Variable	Tipo de Dato	Descripción
Fecha de venta	Fecha	Fecha en la que se registró la transacción de venta.
Año	Texto	Año relacionado a la fecha de venta.
Mes	Texto	Mes relacionado a la fecha de venta.
# Documento	Texto	Identificador único de la factura o documento de venta.
Tipo	Categórica	Tipo de comprobante comercial registrado (Factura, Nota de Crédito, etc.).
RUC	Texto	Identificador único del cliente utilizado para el análisis histórico.
Cliente	Texto	Nombre o razón social del cliente.
Tipo de cliente	Categórica	Clasificación del cliente según el segmento comercial (privado, público, entre otros).
JDL	Categórica	Responsable de la línea de negocio
Código del producto	Texto	Identificador único del producto comercializado.
Producto	Texto	Nombre o descripción del producto vendido.
Categoría	Categórica	Clasificación del producto (reactivos, consumibles, controles, equipos, etc.).
Línea	Categórica	Línea comercial a la que pertenece el producto (Inmunología, Hematología, Química Clínica, entre otras).
Marca	Categórica	Marca de productos comercializados.
Cantidad	Numérica (entera)	Número de unidades vendidas en cada transacción. Variable objetivo de la investigación.
Costo unitario	Numérica (decimal)	Costo asociado a cada unidad del producto.
MG Bruto	Numérica (decimal)	Precio de venta - costo del producto.
Facturación	Numérica (decimal)	Valor total facturado en la transacción de venta.
Ciudad	Categórica	Ciudad donde se encuentra ubicado el cliente.
Provincia	Categórica	Provincia donde se ubica el cliente.
Costo Total	Numérica (decimal)	Costo total de cada producto.

Fuente: Elaboración Propia

3.4. Proceso de Limpieza

Se verificó la calidad de la información mediante la revisión de registros inconsistentes, valores nulos y valores atípicos. Para el tratamiento de estos últimos se utilizó el método del

rango intercuartílico (IQR), identificándose un 4,1 % de outliers, los cuales fueron reemplazados por los límites definidos por este método para reducir su impacto en el entrenamiento de los modelos predictivos. Este procedimiento permitió mantener todos los registros disponibles y minimizar la influencia de valores extremos, mejorando la consistencia y confiabilidad del conjunto de datos utilizado en las etapas posteriores del análisis.

3.5. Preparación de Datos

La fase de preparación de los datos se desarrolló conforme a los lineamientos de la metodología CRISP-ML(Q), con el objetivo de garantizar la calidad y la adecuación de la información para las etapas de segmentación y de modelado predictivo. Durante esta fase se realizaron procesos de transformación de datos, la construcción de las variables RFM (Recency, Frequency y Monetary) y la estandarización de las variables numéricas mediante StandardScaler, obteniendo un conjunto de datos consistente y apto para la aplicación de los algoritmos de clustering y Machine Learning.

3.6. Segmentación de clientes

A partir de los 67.326 registros históricos de ventas, se identificaron 1.824 clientes únicos, sobre los cuales se realizó la agregación de las variables RFM para caracterizar su comportamiento de compra. Estos clientes fueron posteriormente agrupados mediante el algoritmo K-Means en tres clústeres.

La segmentación de clientes se realizó mediante el algoritmo K-Means, utilizando las variables RFM (Recency, Frequency y Monetary), que permiten caracterizar el comportamiento de compra de los clientes. Previamente, las variables fueron estandarizadas para evitar diferencias de escala que pudieran influir en el proceso de agrupamiento. La cantidad adecuada de segmentos se estableció utilizando el método del codo (Elbow Method)

y el coeficiente de Silhouette, lo que permitió identificar tres grupos de clientes con características de compra diferenciadas.

Finalmente, el segmento asignado a cada cliente se incorporó como una nueva variable categórica para su uso en el desarrollo de los modelos predictivos de demanda.

Tabla 2. Distribución clientes por clústers encontrados.

cluster	num_clientes
0	621
1	1.172
2	31
Total	**1.824**

Fuente: Elaboración Propia

3.7. Partición estratificada del dataset

Para el entrenamiento y la evaluación de los modelos predictivos, el conjunto de datos se dividió en un 80 % para entrenamiento y un 20 % para prueba. Esta partición permitió entrenar los algoritmos con una muestra representativa de los datos y, posteriormente, evaluar su capacidad de generalización sobre información no utilizada durante el entrenamiento.

Tabla 3. Distribución de la partición del dataset.

Conjunto	Registros	Porcentaje
Entrenamiento (Train)	53.860	80%
Prueba (Test)	13.466	20%
Total	**67.326**	**100%**

Fuente: Elaboración Propia

3.8. Desarrollo de modelos predictivos

Una vez obtenida la segmentación de clientes, se desarrollaron modelos predictivos de demanda mediante los algoritmos de regresión lineal, Random Forest y XGBoost. Para ello, se usaron las variables Mes, Línea, Categoría y Marca, excluyendo la variable Producto debido a su alta cardinalidad, con aproximadamente 1.455 productos únicos.

La selección de las variables Mes, Línea, Categoría y Marca se realizó considerando su relevancia comercial y su posible influencia en el comportamiento de la demanda, la variable Mes permite captar efectos estacionales, mientras que Línea, Categoría y Marca representan características de los productos asociadas a los patrones de compra. Por su parte, otras variables fueron excluidas por criterios de alta cardinalidad, menor aporte esperado al modelo y validación con el área comercial, priorizando la interpretabilidad y eficiencia de los modelos predictivos.

Adicionalmente, las variables Facturación, Costo Unitario, MG Bruto y Costo Total no fueron utilizadas como variables predictoras debido a que mantienen una dependencia directa con la variable objetivo Cantidad o con los resultados de las transacciones comerciales. Su incorporación habría generado fuga de información (data leakage), al introducir información que no estaría disponible al momento de realizar una predicción futura, afectando la validez de los resultados obtenidos.

Al usar la técnica de One-Hot Encoding su incorporación habría generado un elevado número de variables derivadas durante el proceso de codificación, incrementando la dimensionalidad del conjunto de datos, el costo computacional del entrenamiento y el riesgo de sobreajuste. Por esta razón, se optó por utilizar variables de mayor nivel de agregación, que conservan la información relevante para la predicción de la demanda. Este procedimiento se aplicó de manera consistente en los tres algoritmos, asegurando que todos los modelos fueran entrenados bajo un mismo esquema de preparación de los datos.

Con el propósito de evaluar el aporte de la segmentación de clientes al desempeño predictivo, se plantearon dos escenarios experimentales: el primero, que utilizó únicamente las variables mencionadas anteriormente, y el segundo, que incorporó la variable de clúster generada mediante el algoritmo K-Means. Posteriormente, se comparó el desempeño de los modelos obtenidos en ambos escenarios para determinar la contribución de la segmentación a la precisión de las predicciones.

3.9. Estrategia de configuración y ajuste de modelos

Con la finalidad de optimizar la capacidad predictiva de los modelos, se optimizaron los hiperparámetros. Para los algoritmos Random Forest y XGBoost se empleó la biblioteca Optuna, que permitió explorar y seleccionar de manera automatizada las configuraciones de hiperparámetros más adecuadas, utilizando la raíz del error cuadrático medio (*Root Mean Squared Error*, RMSE) como métrica de evaluación. La validación de cada configuración se realizó mediante TimeSeriesSplit con cinco particiones, lo que garantizó que el procedimiento respetara la naturaleza temporal de los datos y evitara fugas de información entre los conjuntos de entrenamiento y de validación. Por otra parte, para el modelo de Regresión Lineal se utilizaron los parámetros predeterminados de la biblioteca scikit-learn, ya que este algoritmo presenta un número limitado de hiperparámetros susceptibles de optimización.

3.10. Regresión Lineal

Se implementó como modelo base (*baseline*) debido a su simplicidad, su facilidad de interpretación y su amplia utilización en problemas de regresión. Este algoritmo establece una relación lineal entre las variables independientes y la variable objetivo, lo que permite obtener una referencia con la que comparar el desempeño de modelos de mayor complejidad.

Se utilizó la implementación de LinearRegression de la biblioteca scikit-learn, con su configuración estándar, ya que este modelo no requiere ajuste de hiperparámetros. El entrenamiento se realizó utilizando el conjunto de datos previamente dividido en entrenamiento y prueba, en los dos escenarios planteados: con y sin la incorporación de la variable de clúster obtenida mediante K-Means.

Ilustración 1. Código Entrenamiento modelo Regresión Lineal.

```
from sklearn.linear_model import LinearRegression

mlr_model = LinearRegression()
mlr_model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

Fuente: Elaboración Propia

3.11. Random Forest

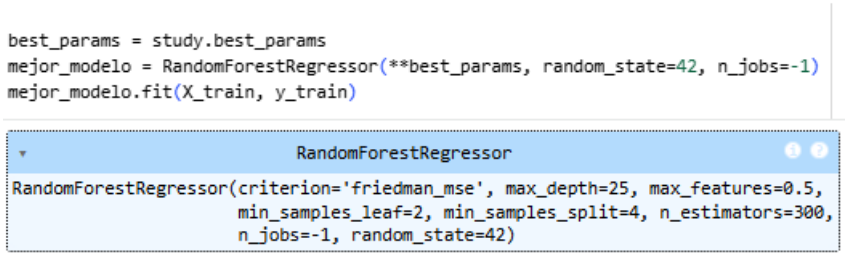
El algoritmo Random Forest fue seleccionado por su capacidad para modelar relaciones no lineales entre variables y reducir el riesgo de sobreajuste mediante la combinación de múltiples árboles de decisión. Esta característica lo convierte en una alternativa robusta para problemas de predicción.

Con el fin de mejorar su desempeño, se realizó un proceso de optimización de hiperparámetros mediante la biblioteca Optuna. Durante esta etapa se evaluaron diferentes configuraciones de parámetros, como el número de árboles (n_estimators), la profundidad máxima de los árboles (max_depth), el número mínimo de observaciones requeridas para dividir un nodo (min_samples_split), el número mínimo de observaciones en los nodos hoja (min_samples_leaf) y el número de variables consideradas en cada división (*max_features*).

La evaluación de cada configuración se realizó mediante TimeSeriesSplit con cinco particiones y el RMSE como métrica de optimización. Finalmente, el modelo se entrenó con

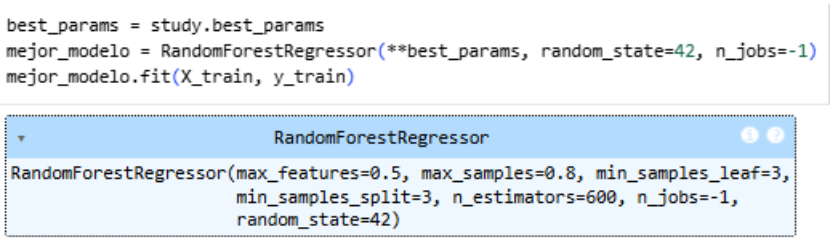
los hiperparámetros óptimos obtenidos y una semilla aleatoria fija (*random_state = 42*) para garantizar la reproducibilidad de los resultados.

Ilustración 2. Código Entrenamiento Random Forest + Clúster.



Fuente: Elaboración Propia

Ilustración 3. Código Entrenamiento Random Forest sin Clúster.



Fuente: Elaboración Propia

3.12. XGBoost

El algoritmo XGBoost (Extreme Gradient Boosting) fue seleccionado por su capacidad para generar modelos predictivos de alta precisión mediante la generación progresiva de árboles de decisión, donde cada nuevo árbol intenta reducir los errores producidos por los modelos anteriores. Esta característica le permite capturar patrones complejos en los datos y alcanzar un alto desempeño en tareas de regresión.

Para optimizar su rendimiento, se utilizó la biblioteca Optuna, que permitió ajustar automáticamente los principales hiperparámetros del modelo. Entre los parámetros evaluados se incluyeron el número de árboles (*n_estimators*), la profundidad máxima de los árboles (*max_depth*), la tasa de aprendizaje (*learning_rate*), la proporción de observaciones utilizadas

en cada iteración (*subsample*), la proporción de variables empleadas en cada árbol (*colsample_bytree*), el peso mínimo requerido para generar nuevas divisiones (*min_child_weight*) y el parámetro de regularización *gamma*.

La evaluación de cada configuración se realizó mediante TimeSeriesSplit con cinco particiones y el RMSE como criterio de optimización. Finalmente, el modelo se entrenó con los hiperparámetros óptimos obtenidos y una semilla aleatoria fija (*random_state = 42*) para garantizar la reproducibilidad de los resultados.

Ilustración 4. Código Entrenamiento XGBoost + Clúster.

```
best_params_xgb = study_xgb.best_params
mejor_modelo_xgb = xgb.XGBRegressor(**best_params_xgb, random_state=42, n_jobs=-1)
mejor_modelo_xgb.fit(X_train, y_train)
```

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.5017576734599645, device=None,
             early_stopping_rounds=None, enable_categorical=True,
             eval_metric=None, feature_types=None, feature_weights=None,
             gamma=3.974624528511253, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.08610337934248824,
             max_bin=None, max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=3, max_leaves=None,
             min_child_weight=3, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=300, n_jobs=-1,
             num_parallel_tree=None, ...)
```

Fuente: Elaboración Propia

Ilustración 5. Código Entrenamiento XGBoost sin Clúster.

```
best_params_xgb = study_xgb.best_params
mejor_modelo_xgb = xgb.XGBRegressor(**best_params_xgb, random_state=42, n_jobs=-1)
mejor_modelo_xgb.fit(X_train, y_train)
```

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.8924849742631324, device=None,
             early_stopping_rounds=None, enable_categorical=True,
             eval_metric=None, feature_types=None, feature_weights=None,
             gamma=2.9721253105037784, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.11403722348087315,
             max_bin=None, max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=3, max_leaves=None,
             min_child_weight=6, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=100, n_jobs=-1,
             num_parallel_tree=None, ...)
```

Fuente: Elaboración Propia

3.13. Evaluación del desempeño de los modelos predictivos

Para evaluar el desempeño de los modelos predictivos desarrollados, se emplearon tres métricas ampliamente utilizadas en problemas de regresión: el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2). Estas métricas permiten analizar la precisión de las predicciones desde diferentes perspectivas y facilitan una comparación objetiva entre los algoritmos evaluados.

- **Error Absoluto Medio (MAE):** Calcula el promedio de las diferencias absolutas entre los valores reales y las estimaciones generadas por el modelo. Al expresarse en las mismas unidades de la variable objetivo, proporciona una medida fácilmente interpretable del error promedio de predicción, valores más bajos de MAE indican un mejor desempeño del modelo.

  Ilustración 6. MAE (Mean Absolute Error).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

  Donde:

  - MAE = 0: Predicción perfecta, no existen errores
  - MAE > 0: el modelo presenta un error promedio de MAE en unidades. A meno
  - MAE: mejor desempeño del modelo.

- **RMSE:** Se obtiene calculando la raíz cuadrada del promedio de los errores cuadráticos, lo que asigna un mayor peso a los errores de gran magnitud. Esta métrica resulta especialmente útil para identificar modelos con menor variabilidad en las predicciones y mayor precisión frente a valores extremos.

Ilustración 7. RMSE (Root Mean Squared Error).

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

Donde:

- RMSE = 0: predicción perfecta, no existen errores.
- RMSE > 0: el modelo presenta un error promedio, penalizando más los errores grandes.
- A menor RMSE: mejor desempeño del modelo. RMSE > MAE: indica la presencia de algunos errores grandes (outliers o predicciones muy alejadas).

- **Coeficiente de determinación (R^2):** Mide el porcentaje de variación de la variable dependiente que puede ser explicado mediante el modelo. Su valor se encuentra entre 0 y 1; mientras más próximo esté a 1, mayor será la capacidad del modelo para explicar los datos y representar adecuadamente su comportamiento.

Ilustración 8. R^2 (Coeficiente de Determinación).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

Donde:

- R^2 = 1: ajuste perfecto.
- R^2 = 0: el modelo no mejora respecto a usar la media.
- R^2 < 0: el modelo es peor que predecir siempre la media

Estas métricas se utilizaron para comparar el desempeño de la Regresión Lineal, Random Forest y XGBoost en los dos escenarios planteados en la investigación (con y sin la

incorporación de la variable Clúster), lo que permitió identificar el modelo con mayor capacidad predictiva para la demanda de productos médicos de laboratorio.

4. Análisis y Discusión de Resultados

4.1. Análisis exploratorio de datos.

Durante esta estapa de entendimiento de los datos, se realizó un análisis exploratorio con el propósito de conocer la estructura, la distribución y la calidad de la información antes del desarrollo de los modelos de segmentación y predicción. Para ello, se analizaron 67.326 registros históricos de ventas correspondientes al período comprendido entre enero de 2023 y marzo de 2026, conformados por 21 variables relacionadas con clientes, productos y transacciones comerciales.

El análisis incluyó la obtención de estadísticas descriptivas, la evaluación de la distribución de las variables numéricas mediante histogramas y diagramas de caja, así como la identificación de valores atípicos y la verificación de la consistencia de la información. Como resultado, se detectó que aproximadamente el 4,1 % de las observaciones correspondía a valores atípicos. Para reducir su influencia en el proceso de modelado, estos fueron tratados mediante el método del rango intercuartílico (IQR), reemplazándolos por los límites establecidos para cada variable, lo que permitió conservar la totalidad de los registros y obtener un conjunto de datos más consistente para las etapas posteriores de segmentación de clientes y desarrollo de los modelos predictivos.

Tabla 4. Estadísticos descriptivos de las variables numéricas.

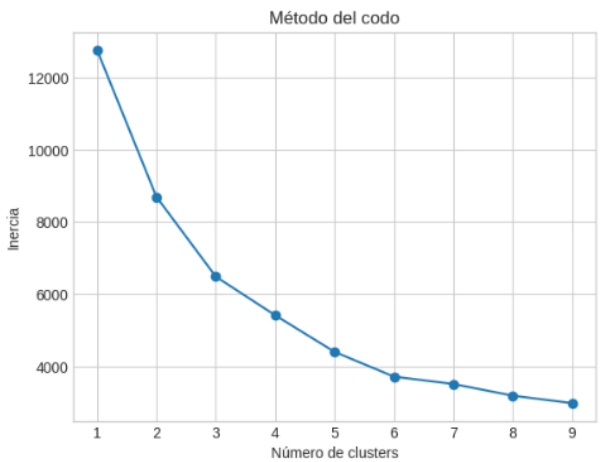
Estadístico	Cantidad	Facturación	Costo Unitario	MG Bruto
count	67.326	67.326	67.326	67.326
mean	2	78	28	49
min	1	0	0	-686
0.25	1	34	13	19
0.5	1	65	20	35
0.75	1	104	33	65
max	252	211	392	528
std	4	53	26	46

Fuente: Elaboración Propia

4.2. Resultados de la segmentación de clientes.

Con el objetivo de identificar grupos de clientes con comportamientos de compra similares, se aplicó el algoritmo K-Means a las variables RFM y a las variables derivadas del comportamiento de compra seleccionadas para la investigación. La segmentación permitió agrupar a los clientes en conjuntos homogéneos.

Ilustración 9. Método del codo (Elbow Method).



Fuente: Elaboración Propia

Ilustración 10. Silhouette Score.

```
K=2 -> Silhouette Score: 0.4117
K=3 -> Silhouette Score: 0.3209
K=4 -> Silhouette Score: 0.2704
K=5 -> Silhouette Score: 0.2861
K=6 -> Silhouette Score: 0.2827
K=7 -> Silhouette Score: 0.2748
K=8 -> Silhouette Score: 0.2604
K=9 -> Silhouette Score: 0.2611
```

Fuente: Elaboración Propia

Para determinar el número óptimo de grupos se empleó el método del codo, complementado con el índice de Silhouette. Aunque el valor del índice de Silhouette para k=3 no fue el máximo, la diferencia respecto a otras configuraciones se redujo.

Considerando el punto de inflexión identificado mediante el método del codo, la adecuada separación entre los grupos y la facilidad de interpretación de los resultados para los objetivos de la investigación, se seleccionó una solución de tres clústeres para el proceso de segmentación.

Adicionalmente, esta configuración fue validada con el área comercial de la empresa, que consideró que la clasificación en tres grupos de clientes resultaba más coherente con la realidad del negocio y facilitaba la interpretación y la aplicación práctica de los resultados en la gestión comercial.

4.3. Resultados de Modelos Predictivos

4.3.1. Regresión Lineal

Los resultados evidencian que la incorporación de la variable Clúster redujo los valores de MAE, RMSE y MAPE, lo que indica una disminución de los errores de predicción respecto al escenario sin segmentación. Este comportamiento sugiere que la información proporcionada

por la segmentación mejoró la precisión de las estimaciones del modelo. Sin embargo, el coeficiente de determinación (R^2) disminuyó, lo que revela una menor capacidad para explicar la variabilidad de la demanda. En este sentido, la variable clúster contribuyó a reducir los errores de predicción, pero no mejoró el ajuste global del modelo.

Tabla 5. Resultados del modelo de Regresión Lineal.

Regresión Lineal	MAE	RMSE	R2	MAPE (%)
Con Clúster	4.8916	16.9028	0.2557	153.96%
Sin Clúster	8.89	24.2659	0.3727	227.85%

Fuente: Elaboración Propia

4.3.2. Resultados de Random Forest

Los resultados evidencian que la incorporación de la variable clúster redujo los valores de MAE, RMSE y MAPE, lo que indica una disminución de los errores de predicción respecto al escenario sin segmentación. Este comportamiento sugiere que la información derivada de la segmentación contribuyó a aumentar la precisión de las estimaciones del modelo Random Forest. No obstante, el coeficiente de determinación (R^2) disminuyó, lo que indica una menor capacidad para explicar la variabilidad de la demanda. En consecuencia, aunque la variable clúster permitió reducir los errores de predicción, no mejoró el ajuste global del modelo.

Tabla 6. Resultados de Random Forest.

Random Forest	MAE	RMSE	R2	MAPE (%)
Con Clúster	4.8362	16.518	0.2892	143.47%
Sin Clúster	8.4694	22.8642	0.443	201.32%

Fuente: Elaboración Propia

4.3.3. Resultados de XGBoost

Los resultados evidencian que la incorporación de la variable clúster redujo los valores de MAE, RMSE y MAPE, lo que indica una disminución de los errores de predicción respecto al escenario sin segmentación. Este comportamiento sugiere que la información derivada de la segmentación contribuyó a mejorar la precisión de las estimaciones del modelo XGBoost. No obstante, el coeficiente de determinación (R^2) disminuyó, lo que indica una menor capacidad para explicar la variabilidad de la demanda. En consecuencia, aunque la variable clúster permitió reducir los errores de predicción, no mejoró el ajuste global del modelo.

Tabla 7. Resultados de XGBoost.

XGBoost	MAE	RMSE	R2	MAPE (%)
Con Clúster	4.9384	16.2947	0.3083	155.77%
Sin Clúster	8.8188	22.6078	0.4555	228.21%

Fuente: Elaboración Propia

4.4. Comparación de los modelos predictivos

Para determinar el modelo con mejor desempeño y evaluar el efecto de incorporar la variable de clúster, se compararon los resultados obtenidos por los algoritmos de Regresión Lineal, Random Forest y XGBoost en los dos escenarios planteados en la investigación. La comparación se realizó mediante las métricas MAE, RMSE, R^2 y MAPE, lo que permitió analizar las diferencias en la precisión de las predicciones.

Tabla 8. Comparación del desempeño de los modelos predictivos.

Modelo	Escenario	MAE	RMSE	R^2	MAPE (%)
Regresión Lineal	Con Clúster	4.8916	16.9028	0.2557	153.96%
Regresión Lineal	Sin Clúster	8.89	24.2659	0.3727	227.85%
Random Forest	Con Clúster	4.8362	16.518	0.2892	143.47%
Random Forest	Sin Clúster	8.4694	22.8642	0.4430	201.32%
XGBoost	Con Clúster	4.9384	16.2947	0.3083	155.77%
XGBoost	Sin Clúster	8.8188	22.6078	0.4555	228.21%

Fuente: Elaboración Propia

En general, los resultados muestran que la incorporación de la variable Clúster permitió reducir los valores de MAE, RMSE y MAPE en los tres algoritmos evaluados. No obstante, el R^2 disminuyó en todos los modelos, lo que indica que la segmentación no mejoró su capacidad explicativa. Este comportamiento podría estar relacionado, entre otros factores, con el marcado desbalance de los grupos obtenidos, particularmente por la baja representación del clúster 2.

Al comparar el desempeño de los algoritmos, Random Forest obtuvo los valores más bajos de MAE y MAPE en el escenario con segmentación, mientras que XGBoost presentó el menor valor de RMSE. Sin embargo, considerando el criterio de selección establecido en esta investigación, basado principalmente en la capacidad explicativa del modelo (R^2), el mejor desempeño se obtuvo con XGBoost sin incorporar la variable Clúster, con un R^2 de 0.4555.

Estos resultados sugieren que, aunque la segmentación de clientes constituye un recurso valioso para caracterizar el comportamiento de compra y apoyar las estrategias comerciales, su incorporación como variable explicativa no mejoró la capacidad predictiva de los modelos diseñados para pronosticar la demanda. En consecuencia, para la predicción de la demanda de

productos médicos de laboratorio, el modelo XGBoost sin la variable Clúster se presenta como la alternativa con el mejor desempeño en el contexto de esta investigación.

4.5. Discusión de Resultados

Los resultados obtenidos evidencian que la segmentación de clientes mediante el algoritmo K-Means permitió identificar grupos con comportamientos de compra diferenciados, lo cual coincide con investigaciones previas que destacan la utilidad del clustering para comprender los patrones de consumo y apoyar la toma de decisiones comerciales. En este sentido, los hallazgos son consistentes con lo reportado por Pincay (2023), Rojo (2024) y Dunaboina (2024), quienes concluyen que la segmentación facilita la identificación de perfiles de clientes con características similares y constituye una herramienta de apoyo para el diseño de estrategias comerciales y de fidelización.

Respecto a los modelos predictivos, los tres algoritmos evaluados presentaron un comportamiento similar al incorporar la variable Clúster. En todos los casos se observó una reducción de las métricas de error (MAE, RMSE y MAPE); sin embargo, el coeficiente de determinación (R^2) disminuyó en comparación con los modelos desarrollados sin dicha variable. Estos resultados indican que, para el conjunto de datos analizado, la segmentación de clientes no incrementó la capacidad explicativa de los modelos de predicción de demanda, aunque sí contribuyó a disminuir los errores absolutos de predicción.

Un aspecto importante es el desbalance observado en la distribución de los clústeres. El clúster 2 concentra únicamente 31 clientes (1,7 % del total), mientras que los clústeres 0 y 1 agrupan el 34,0 % y 64,3 %, respectivamente. Al incorporar esta variable mediante One-Hot Encoding, la categoría minoritaria dispone de una cantidad reducida de observaciones, lo que podría limitar su aporte al aprendizaje de los modelos e introducir variabilidad adicional. Este

comportamiento podría explicar, en parte, la reducción del R^2 observada al incorporar la variable Clúster.

En la comparación entre algoritmos, XGBoost obtuvo el mayor valor de R^2 (0.4555) en el escenario sin la incorporación de la variable Clúster, mientras que Random Forest presentó los menores valores de MAE y MAPE. Este comportamiento coincide con estudios como los de Nasseri et al. (2023) y Saranani et al. (2024), quienes reportan un desempeño superior de los algoritmos basados en árboles de decisión y técnicas de boosting frente a modelos lineales en tareas de predicción de demanda. En contraste, la Regresión Lineal mostró un desempeño inferior, lo cual sugiere que las relaciones presentes en los datos analizados no pueden ser representadas adecuadamente mediante un modelo estrictamente lineal.

Finalmente, uno de los principales aportes de esta investigación radica en evaluar la incorporación de la segmentación de clientes como variable explicativa dentro de modelos predictivos de demanda. A diferencia de estudios que integran ambas metodologías y reportan mejoras en el desempeño predictivo, como Kasem et al. (2024) y Ariwiati et al. (2026), en esta investigación dicha incorporación no produjo un incremento en la capacidad explicativa de los modelos. Este resultado evidencia que la utilidad de la segmentación depende del contexto, de las características de los datos y del problema analizado. En consecuencia, para el caso de estudio desarrollado, la segmentación constituye una herramienta valiosa para caracterizar el comportamiento de los clientes, mientras que el modelo XGBoost sin la variable Clúster presentó las mejores condiciones para realizar el pronóstico de la demanda de productos médicos de laboratorio.

5. Conclusiones y Recomendaciones

5.1. Conclusiones

La presente investigación permitió demostrar la viabilidad de aplicar técnicas de Machine Learning para la predicción de la demanda de productos médicos de laboratorio, integrando procesos de segmentación de clientes y modelos predictivos en el marco metodológico CRISP-ML(Q). La metodología proporcionó un proceso sistemático para el desarrollo del proyecto, lo que permitió transformar los registros históricos de ventas en información útil para apoyar la toma de decisiones comerciales basadas en datos.

Los resultados del análisis exploratorio permitieron caracterizar el comportamiento de los datos históricos de ventas del período comprendido entre enero de 2023 y marzo de 2026. Asimismo, el proceso de preparación de datos garantizó la calidad de la información mediante la depuración de registros, el tratamiento de valores atípicos mediante el método del rango intercuartílico (IQR) y la transformación de las variables requeridas para el desarrollo de los modelos analíticos.

En la etapa de segmentación se identificaron grupos de clientes con comportamientos de compra diferenciados mediante el algoritmo K-Means, lo que evidencia que la segmentación es una herramienta útil para caracterizar los distintos perfiles de clientes y comprender sus patrones de consumo. Sin embargo, al incorporar la variable Cluster en los modelos predictivos, se observó que, aunque las métricas de error (MAE, RMSE y MAPE) disminuyeron en los tres algoritmos evaluados, el coeficiente de determinación (R^2) disminuyó respecto al escenario sin segmentación. Esto indica que, para el conjunto de datos analizado, la información derivada de la segmentación no incrementó la capacidad explicativa de los modelos para predecir la demanda.

La comparación entre los algoritmos desarrollados permitió concluir que los modelos basados en árboles de decisión presentaron un mejor desempeño que la regresión Lineal. En particular, XGBoost obtuvo el mayor coeficiente de determinación ($R^2 = 0.4555$) en el escenario sin la variable Cluster, por lo que fue seleccionado como el modelo con la mejor capacidad predictiva para estimar la demanda de productos médicos de laboratorio. Este resultado evidencia que el uso de algoritmos de ensamble constituye una alternativa adecuada para modelar las relaciones complejas presentes en los datos de ventas.

Finalmente, la investigación aporta evidencia empírica para el sector de productos médicos de laboratorio en el Ecuador, demostrando que la segmentación de clientes constituye una herramienta valiosa para el análisis del comportamiento comercial, aunque su incorporación como variable explicativa no necesariamente mejora el desempeño de los modelos predictivos. En este sentido, los resultados obtenidos contribuyen al desarrollo de soluciones analíticas orientadas a fortalecer la toma de decisiones comerciales mediante el uso de técnicas de Ciencia de Datos y Machine Learning.

5.2. Recomendaciones

Se recomienda implementar el modelo XGBoost como herramienta de apoyo para la predicción de la demanda de productos médicos de laboratorio, integrándolo en los procesos de planificación comercial y de abastecimiento, con el fin de fortalecer la toma de decisiones basada en datos.

Se sugiere mantener un proceso continuo de actualización del modelo mediante la incorporación periódica de nuevos registros históricos de ventas, de modo que las predicciones reflejen los cambios en el comportamiento de los clientes y en las condiciones del mercado.

Resulta conveniente ampliar el conjunto de variables utilizadas en futuras investigaciones mediante la incorporación de factores externos, como indicadores económicos, la estacionalidad, las campañas comerciales o variables relacionadas con la disponibilidad de productos, con el propósito de incrementar la capacidad predictiva de los modelos desarrollados.

Asimismo, se recomienda utilizar la segmentación de clientes como herramienta de apoyo para el diseño de estrategias de marketing, fidelización y gestión comercial, aprovechando la información obtenida sobre los diferentes perfiles de clientes, aun cuando su incorporación no haya mejorado la capacidad predictiva de los modelos de demanda en este estudio.

Finalmente, se propone que futuras investigaciones evalúen otros algoritmos de Machine Learning y técnicas de optimización de hiperparámetros, así como conjuntos de datos con mayor cobertura temporal y geográfica, con el objetivo de comparar su desempeño y determinar nuevas alternativas para mejorar la precisión en la predicción de la demanda.

Referencias

Aldahmani, E., Alzubi, A., & Iyiola, K. (2024). Demand Forecasting in Supply Chain Using Uni-Regression Deep Approximate Forecasting Model. *Applied Sciences*, *14*(18), 8110. https://doi.org/10.3390/app14188110

Alonso, A. M. (2024). *Desarrollo de Modelos de Pronóstico de Demanda con Variables Predictoras para Mejorar la Gestión de Inventarios en una Empresa de Alimentos* [Tesis de Pregrado, Universidad de Chile]. https://repositorio.uchile.cl/bitstream/handle/2250/204012/Desarrollo-de-modelos-de-pronostico-de-demanda-con-variables-predictoras-para-mejorar-la-gestion-de-inventarios-en-una-empresa-de-alimentos.pdf;jsessionid=5FE464311C46E843B7AC7AD7DBDC5F6F?sequence=1

Ariwiati, Susanty, A., & Batu, K. L. (2026). Product Category Level Demand Forecasting Based on LRFM–CLV Customer Segmentation in B2B E-Marketplace Systems. *Engineering, Technology & Applied Science Research*, *16*(3), 35205-35212. https://doi.org/10.48084/etasr.18169

Mankiw, N. G. (2021). Principles of economics (9th ed.). Cengage Learning.

Becerra, A. (2024). *Predicción de ventas en tiendas de retail farmacéutico ecuatoriano mediante modelos agrupados de aprendizaje automático y series temporales* [Tesis de Posgrado, Pontificia Universidad Católica del Ecuador]. https://repositorio.puce.edu.ec/server/api/core/bitstreams/a5129e40-139f-429d-bf48-5db7f213142f/content?utm_source=chatgpt.com

Calle, A. J., Aviles, E. M., Baque, E. A., & Muñiz, F. S. (2024). El papel de la analítica predictiva en la anticipación de cambios en el entorno empresarial. *Ciencia y Desarrollo*, *27*(2), 43. https://doi.org/10.21503/cyd.v27i2.2601

Castro, J. R., Mendoza, M. A., & Soberanes, A. (2025). Algoritmos de clustering y técnicas de IA para la identificación de patrones espaciales: Una revisión sistemática. *Religación*, *10*(48), e2501570. https://doi.org/10.46652/rgn.v10i48.1570

Abdul-Azeez, O., Ihechere, A. O., & Idemudia, C. (2024). Enhancing business performance: The role of data-driven analytics in strategic decision-making. International Journal of Management & Entrepreneurship Research, 6(7), 2066-2081.

Dias, B. L. (2025). Healthcare Supply Chain Analytics: Smart Supply Chain AI Powered Inventory Optimization in Healthcare. *Journal of Data Analysis and Critical Management*, *1*(01). https://doi.org/10.64235/82060w24

Diez, A. A. (2022). *Segmentación RFM de clientes de un programa de fidelización en una empresa retail con el algoritmo de análisis clúster PAM* [Tesis de Pregrado, Universidad Nacional Agraria La Molina]. https://repositorio.lamolina.edu.pe/server/api/core/bitstreams/a46fd122-c508-4184-b36c-a7b46cf77765/content

Dunaboina, B. (2024). *Customer Segmentation Approach for Efficient Targeting in Retail* [Master Thesis, Dublin Business School]. https://esource.dbs.ie/server/api/core/bitstreams/80c52948-bbf4-4e72-aafc-9b812623ac76/content

Eglite, L., & Birzniece, I. (2022). Retail Sales Forecasting Using Deep Learning: Systematic Literature Review. *Complex Systems Informatics and Modeling Quarterly*, (30), 53-62. https://doi.org/10.7250/csimq.2022-30.03

Forero, W., & Negre, F. (2023). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: Una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia*, *27*(1), 209-253. https://doi.org/10.5944/ried.27.1.37491

González, Á. A. F. (2021). *Machine learning para la segmentación y optimización de los costos de adquisición de clientes* [Tesis de Posgrado, Los Libertadores Funación Universitaria]. https://repository.libertadores.edu.co/server/api/core/bitstreams/314bcc9e-8b95-46c5-99ec-81f07b466da4/content

Hanaysha, J. R., Al Shaikh, M. E., & Alzoubi, H. M. (2021). Importance of Marketing Mix Elements in Determining Consumer Purchase Decision in the Retail Market: *International Journal of Service Science, Management, Engineering, and Technology*, *12*(6), 56-72. https://doi.org/10.4018/IJSSMET.2021110104

Hermitaño, J. A. (2022). Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática. *Interfases*, (015), 160-178. https://doi.org/10.26439/interfases2022.n015.5898

Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995-5005. https://doi.org/10.1007/s00521-023-09339-6

León, S. M. (2025). *Modelos de Predicción en Machine Learning: Una Revisión* [Tesis de Pregrado, Universidad de La Laguna]. https://riull.ull.es/xmlui/bitstream/handle/915/42994/Modelos%20de%20prediccion%20en%20Machine%20Learning%20Una%20revision.pdf?sequence=1&isAllowed=y

López, S. (2022). *Segmentación de clientes con afectación en sus servicios Área Analytics TIGO* [Tesis de Pregrado, Universidad de Antioquia]. https://hdl.handle.net/10495/29116

Mariño, M. D., Arango, A., Lotero, L., & Jimenez, M. (2021). Modelos de series temporales para pronóstico de la demanda eléctrica del sector de explotación de minas y canteras en Colombia. *Revista EIA*, *18*(35). https://doi.org/10.24050/reia.v18i35.1458

Marrero, L., Carrizo, D., García-Santander, L., & Ulloa, F. (2021). Uso de algoritmo K-means para clasificar perfiles de clientes con datos de medidores inteligentes de consumo eléctrico: Un caso de estudio. *Ingeniare. Revista Chilena de Ingeniería*, *29*(4), 778-787. https://doi.org/10.4067/S0718-33052021000400778

Mellet, D., Alva, D., & Vilchez, D. (2026). *Propuesta de modelo de Inteligencia de Clientes basado en RFM, Big Data y Machine Learning para la fidelización del cliente en Monark* [Tesis de Posgrado, Universidad Peruana de Ciencias Aplicadas (UPC)]. https://repositorioacademico.upc.edu.pe/server/api/core/bitstreams/eac451a2-b12a-5f01-99c9-fbe2999abb6e/content

Mohammad, N. (2023). *A Computational Theory and Semi-Supervised Algorithm for Clustering* (Versión 2). arXiv. https://doi.org/10.48550/ARXIV.2306.06974

Muñoz, J. P., Cárdenas, M., & Heredia, J. (2022). Segmentación de clientes residenciales con fines de autoabastecimiento fotovoltaico mediante índices de rentabilidad en Ecuador. *I+D Tecnológico*, *18*(2), 32-43. https://doi.org/10.33412/idt.v18.2.3646

Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying Machine Learning in Retail Demand Prediction—A Comparison of Tree-Based Ensembles and Long Short-Term Memory-Based Deep Learning. *Applied Sciences*, *13*(19), 11112. https://doi.org/10.3390/app131911112

Nivela, M. A. (2024). Modelo predictivo usando Machine Learning para determinar millas por galón. *Código Científico Revista de Investigación*, *5*(1), 1528-1545. https://doi.org/10.55813/gaea/ccri/v5/n1/467

Pertuz, C. M. P. (2022). *Aprendizaje automático y profundo en Python*. Ra-Ma Editorial.

Pincay, J. V. (2023). *Minería de Datos para Segmentación de clientes en el Laboratorio Clínico particular Pura Vida* [Tesis de Pregrado, Pontificia Universidad Católica del Ecuador]. https://repositorio.puce.edu.ec/server/api/core/bitstreams/c7c45825-f872-47d3-9f9d-00d0eeecbaf5/content

Prakash, D. (2024). Data-Driven Management: The Impact of Big Data Analytics on Organizational Performance. *International Journal for Global Academic & Scientific Research*, *3*(2), 12-23. https://doi.org/10.55938/ijgasr.v3i2.74

Ranabhatt, N., Barreto, S., Pimpão, M., & Prates, P. (2025). Demand Forecasting in the Automotive Industry: A Systematic Literature Review. *Forecasting*, *7*(4), 73. https://doi.org/10.3390/forecast7040073

Rojo, S. (2024). *Segmentación de clientes y optimización de su fidelización mediante aprendizaje computacional* [Tesis de Posgrado, Universidad Oberta de Catalunya]. https://openaccess.uoc.edu/items/b3964d46-2263-45ff-80f7-5205722ec361#page=1

Saranani, F., Dahiya, R., Geeta, S. D. T., Khan, P. H., & Nagina, R. (2024). Pharmaceutical Sales Forecasting with Machine Learning: A Strategic Management Tool for Decision-Making. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(21s), 907-914. https://doi.org/ijisae.org/index.php/IJISAE/article/view/5488

Seyedan, S. (2023). *Development of Predictive Analytics for Demand Forecasting and Inventory Management in Supply Chain using Machine Learning Techniques* [Doctoral Thesis, Concordia Institute for Information Systems Engineering (CIISE)]. https://spectrum.library.concordia.ca/id/eprint/993482/1/Seyedan_PhD_S2024.pdf

Singh, H. V., Girdhar, A., & Dahiya, S. (2022). A Literature survey based on DBSCAN algorithms. *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 751-758. https://doi.org/10.1109/ICICCS53718.2022.9788440

Sodiq, A., & Joy, O. (2024). Leveraging Predictive Analytics to Improve Demand Forecasting and Inventory Management in Healthcare Supply Chains. *International Journal of Scientific Research in Science, Engineering and Technology*, *11*(2), 624-644. https://doi.org/10.32628/IJSRSET2512184

Syamril, S., Herman, H., & Arsyad, M. (2024). Analysis of Factors Affecting Consumer Demand. *Golden Ratio of Marketing and Applied Psychology of Business*, *4*(2), 220-228. https://doi.org/10.52970/grmapb.v4i2.687

Vichi, M., Cavicchia, C., & Groenen, P. J. F. (2022). Hierarchical Means Clustering. *Journal of Classification*, *39*(3), 553-577. https://doi.org/10.1007/s00357-022-09419-7

Wu, Q. (2023). Research on the Factors that Affect Demand and the Effects of Changes in Demand on Prices. *Advances in Economics, Management and Political Sciences*, *18*(1), 391-395. https://doi.org/10.54254/2754-1169/18/20230104

Mitchell, M. (1997). Machine learning. McGraw-Hill.

Studer, S., Benjamins, R., Bongard, J., Knobloch, M., Kurpicz-Briki, M., Reichstein, M., ... & Huber, S. (2021). Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology. Machine Learning and Knowledge Extraction, 3(2), 392-413. https://doi.org/10.3390/make3020020

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Waltham, MA: Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

Thiyagarajan, S., Kumar, R., & colaboradores. (2026). Hybrid clustering-based demand forecasting models: A comparative analysis. [Verificar datos completos de publicación].