FACULTAD DE INGENIERÍA
INGENIERÍA EN COMPUTACIÓN E INFORMÁTICA

ESTUDIO DE PATRONES DE COMPRA Y SEGMENTACIÓN DE CLIENTES EN SUPERMERCADOS ONLINE MEDIANTE MINERÍA DE DATOS: CASO LIDER.CL

Proyecto de título para optar al título de Ingeniero en Computación e Informática

Autor
ALFONSO IGNACIO ALEJANDRO ABBOTT VIDAL

PROFESOR GUÍA
FELIX GONZALO BURGOS GONZALEZ

SANTIAGO, CHILE
2025

FACULTAD DE INGENIERÍA
INGENIERÍA EN COMPUTACIÓN E INFORMÁTICA

DECLARACIÓN DE ORIGINALIDAD Y PROPIEDAD

Yo, **Alfonso Abbott Vidal**, declaro por este medio que el trabajo de titulación presentado para su defensa y evaluación es original; las fuentes, herramientas y aplicaciones utilizadas que contribuyeron a la investigación realizada están debidamente citadas en el texto y acreditadas en el apartado de las referencias, conforme con los requisitos que establece el estilo bibliográfico APA 7.0 y respetando los aspectos que conciernen a la propiedad intelectual.

Por lo tanto, ante cualquier falta de integridad académica encontrada y que atente contra la Ley N°17.336 de Propiedad Intelectual, se asume la responsabilidad que representa para tal efecto, dejando constancia de ello, con fecha 30 de abril de 2025, en la ciudad de Santiago.

Facultad de Ingeniería

Escuela Computación e Informática

Título del trabajo Estudio de patrones de compra y segmentación de clientes en supermercados online mediante minería de datos: caso Lider.cl

Nombre y firma del autor

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO I: INTRODUCCIÓN

1- Introducción

La creciente digitalización del comercio minorista ha generado un aumento significativo en la recolección de datos transaccionales, permitiendo a las empresas conocer con mayor profundidad los hábitos de consumo de sus clientes. En particular, el sector de los supermercados online ha evolucionado rápidamente, impulsado por factores como la comodidad, la inmediatez en las entregas, y la reciente consolidación de modelos de venta híbridos. En Chile, este fenómeno ha sido respaldado por cifras de la Cámara de Comercio de Santiago, que indican que las ventas online superaron los 12.000 millones de dólares en 2023, representando más del 10% del retail nacional (CCS, 2024). Este contexto ofrece oportunidades clave para aplicar técnicas de análisis de datos que potencien la personalización de la oferta, el diseño de campañas promocionales efectivas y la fidelización de los consumidores.

El presente proyecto propone la aplicación de técnicas de minería de datos para estudiar patrones de compra y segmentar clientes en un entorno simulado, tomando como base un conjunto de datos públicos del mercado norteamericano. A través de herramientas como Python, se implementan dos enfoques complementarios: el análisis de reglas de asociación mediante el algoritmo Apriori y la segmentación de clientes utilizando el método de K-Means. Ambos métodos permiten generar conocimiento valioso a partir de grandes volúmenes de información, extrayendo relaciones frecuentes entre productos y agrupando a los clientes según similitudes en sus comportamientos de compra.

El informe se estructura en cuatro capítulos. El Capítulo I entrega una contextualización general del problema y define los objetivos del proyecto, junto con

las métricas, supuestos y alcances. El Capítulo II describe la metodología seguida, detallando las etapas de recopilación, limpieza, transformación y modelamiento de los datos. En el Capítulo III se desarrollan los resultados obtenidos, incluyendo visualizaciones interactivas con Dash y Plotly, y un enfoque de trabajo reproducible en entorno virtual. Finalmente, el Capítulo IV presenta los hallazgos preliminares del análisis: se identifican clústeres con perfiles de consumo diferenciados y reglas de asociación útiles para estrategias de venta cruzada, aportando evidencia para futuras decisiones comerciales.

Este estudio representa una aproximación práctica al uso de técnicas de ciencia de datos en el comercio digital, demostrando cómo el análisis exploratorio, el modelamiento y la visualización pueden ser articulados en una herramienta interactiva para facilitar la toma de decisiones estratégicas en empresas del sector retail.

2- Importancia de resolver el problema

En el contexto actual del comercio digital, las empresas que operan plataformas de venta online se enfrentan a un entorno altamente competitivo, en el cual el conocimiento profundo del comportamiento de sus clientes se convierte en un factor clave para el éxito. Supermercados como Lider.cl, que manejan un amplio catálogo de productos y un volumen masivo de transacciones, requieren estrategias inteligentes para diferenciarse, fidelizar a sus clientes y maximizar sus ingresos. Sin embargo, muchas decisiones comerciales aún se toman sin considerar todo el potencial de la información disponible en los datos transaccionales.

La capacidad de detectar patrones de comportamiento en las compras permite responder preguntas como: ¿qué productos suelen comprarse juntos?, ¿existen segmentos de clientes con preferencias similares?, ¿en qué momentos se concentran ciertas categorías de productos?, entre otras. Resolver estas interrogantes no solo tiene valor comercial, sino que también permite mejorar la experiencia del cliente al ofrecer recomendaciones más precisas, promociones más efectivas y una disposición de productos más alineada con los hábitos reales de consumo.

No contar con estas herramientas analíticas genera desventajas competitivas. Empresas que no logran identificar los patrones de comportamiento de sus clientes tienen mayores probabilidades de diseñar campañas genéricas, desperdiciar recursos publicitarios y perder oportunidades de ventas cruzadas. Además, el desconocimiento de la segmentación real de su clientela limita la posibilidad de crear relaciones personalizadas y sostenidas en el tiempo.

En el caso de supermercados online como Lider.cl, donde la oferta de productos es extensa y la frecuencia de compra es alta, la ausencia de análisis automatizado de

patrones puede derivar en ineficiencias logísticas, sobreoferta de productos poco demandados y pérdida de fidelización en segmentos clave. Por tanto, implementar mecanismos de análisis que permitan explotar los datos disponibles se vuelve una necesidad estratégica.

Este proyecto propone justamente abordar esta necesidad, desarrollando un análisis basado en técnicas de minería de datos orientadas a descubrir asociaciones relevantes entre productos comprados conjuntamente y a segmentar a los clientes en grupos homogéneos. Estas estrategias permitirán apoyar decisiones clave como la ubicación de productos en la plataforma, la generación de recomendaciones automáticas y la planificación de campañas promocionales más efectivas. Resolver este problema contribuirá a mejorar los indicadores clave de rendimiento comercial, como la tasa de conversión, el valor del carrito promedio y la recurrencia de compra.

3- Breve discusión bibliográfica

La minería de datos aplicada a la identificación de patrones de compra y la segmentación de clientes ha sido ampliamente estudiada en la literatura reciente, consolidándose como una herramienta estratégica en el ámbito del comercio electrónico. Diversos enfoques y técnicas han sido propuestos para optimizar la gestión de ventas, la planificación de campañas de marketing y la personalización de la experiencia del cliente.

Agrawal, Imieliński y Swami (1993) sentaron las bases de las reglas de asociación mediante el desarrollo del algoritmo Apriori, una técnica que permite descubrir relaciones significativas entre ítems comprados de manera conjunta en grandes bases de datos. Este método ha sido ampliamente utilizado en la industria para generar estrategias de ventas cruzadas, aumentar el ticket promedio y mejorar la disposición de productos tanto en tiendas físicas como online.

En investigaciones más recientes, Grover, Kar y Ilavarasan (2020) analizaron el uso de técnicas de minería de datos en el sector del retail, destacando la importancia de la personalización basada en patrones de compra para incrementar la fidelización de clientes y la eficiencia en campañas de marketing. Su estudio enfatiza que la identificación de combinaciones frecuentes de productos y la segmentación de clientes son componentes fundamentales para la competitividad en mercados digitales.

Un caso de aplicación práctica es presentado por Song y Kim (2022), quienes implementaron un sistema de recomendación en una plataforma de e-commerce coreana utilizando reglas de asociación y clustering de clientes. Su investigación demostró que la combinación de ambas técnicas permitió mejorar en un 25% la tasa de conversión de usuarios, validando así el impacto positivo de aplicar minería de datos en entornos comerciales reales.

Por otra parte, Ramesh y Baskar (2021) realizaron un estudio comparativo entre diferentes técnicas de clustering aplicadas al comportamiento de compra de clientes, concluyendo que el algoritmo K-means resulta altamente efectivo para agrupar consumidores en segmentos homogéneos que facilitan la planificación de estrategias de marketing diferenciadas.

Finalmente, en el contexto latinoamericano, González et al. (2023) exploraron el potencial del análisis de datos en supermercados online de Chile, destacando la necesidad de integrar procesos de minería de datos como parte de la transformación digital de las empresas del rubro. Su investigación plantea que, pese al crecimiento del e-commerce, aún existe una brecha significativa en el uso de análisis avanzado de datos en empresas del país, lo que representa una oportunidad estratégica para quienes adopten estas tecnologías.

En conjunto, estos estudios sustentan la relevancia y aplicabilidad de los enfoques propuestos en el presente proyecto, validando la elección de técnicas de reglas de asociación y clustering para el análisis de patrones de compra en plataformas de venta online como Lider.cl.

4- Contribución del trabajo

La contribución principal de este proyecto radica en el desarrollo de un análisis de patrones de compra aplicable a supermercados online, específicamente en el contexto de **Lider.cl**, mediante la utilización de técnicas de minería de datos. Esta contribución se materializa en varios niveles:

En primer lugar, se pretende generar conocimiento estratégico a partir de datos transaccionales, facilitando la identificación de productos que son comúnmente adquiridos de manera conjunta. A través de la aplicación de reglas de asociación, se podrán detectar combinaciones frecuentes de artículos, lo que permitirá diseñar estrategias de ventas cruzadas, paquetes promocionales y campañas publicitarias más efectivas, alineadas con el comportamiento real de los clientes.

En segundo lugar, mediante la segmentación de clientes basada en técnicas de clustering, se busca agrupar a los consumidores en perfiles de comportamiento homogéneos. Esto proporcionará a la empresa una comprensión más profunda de sus distintas audiencias, permitiendo personalizar las acciones de marketing, optimizar la gestión del inventario y mejorar la experiencia de compra.

Adicionalmente, este proyecto contribuye al fortalecimiento de la cultura organizacional orientada a la toma de decisiones basada en datos. Al demostrar el valor práctico de la minería de datos aplicada al entorno del e-commerce de

supermercados, se promueve la integración de procesos analíticos avanzados como parte de la estrategia comercial de las empresas chilenas.

Desde el punto de vista tecnológico, el proyecto contempla la implementación de los análisis utilizando herramientas de programación como Python y R, favoreciendo el desarrollo de soluciones reproducibles, escalables y fácilmente adaptables a otros entornos de venta online.

La elección de los algoritmos Apriori y K-means responde a su eficacia comprobada en contextos de retail online. Apriori, desarrollado por Agrawal y Srikant (1994), ha sido ampliamente aplicado para descubrir relaciones de compra frecuentes, mientras que K-means se ha consolidado como una herramienta clave en la segmentación de clientes por su simplicidad y escalabilidad. El uso de Python y R se justifica no solo por su versatilidad, sino también por la disponibilidad de librerías optimizadas para estos algoritmos, como mlxtend, pandas, scikit-learn y ggplot2, lo que facilita la implementación práctica y el análisis exploratorio profundo.

Finalmente, se deja abierta la posibilidad de que los resultados de este análisis puedan, en etapas futuras, integrarse en sistemas de visualización o recomendación en plataformas web, ampliando así el alcance práctico y estratégico del trabajo.

5- Trabajo a realizar en el proyecto

El proyecto contempla el desarrollo de un análisis de datos transaccionales provenientes de un dataset público representativo de ventas online, simulando su aplicación en el contexto del supermercado Lider.cl. El objetivo principal es identificar patrones de comportamiento de compra y segmentar clientes para optimizar las estrategias de venta.

Las actividades principales a realizar en el proyecto incluyen:

- Recopilación y preprocesamiento de datos: selección y limpieza de un conjunto de datos de ventas online que contenga información de productos adquiridos, identificadores de clientes y fechas de transacción.

- Aplicación de reglas de asociación: implementación del algoritmo Apriori u otras técnicas relacionadas para identificar combinaciones frecuentes de productos comprados conjuntamente.

- Segmentación de clientes mediante clustering: aplicación de algoritmos de agrupamiento como K-means u otros métodos adecuados para formar grupos de clientes basados en su comportamiento de compra.

- Análisis e interpretación de resultados: evaluación de los patrones de compra identificados y de las características de los distintos segmentos de clientes, generando insights accionables para estrategias comerciales.

- Documentación de hallazgos: elaboración de reportes que presenten los resultados obtenidos mediante gráficos, tablas y descripciones analíticas claras, facilitando su interpretación para un público no técnico.

Adicionalmente, se plantea como una posibilidad futura el desarrollo de una interfaz web simple para la visualización de los resultados, dependiendo del alcance final del proyecto y la disponibilidad de tiempo.

El proyecto utilizará como herramientas principales los lenguajes de programación Python y R, aprovechando librerías especializadas en análisis de datos y minería de datos como mlxtend, scikit-learn, y arules, entre otras.

CAPÍTULO II: IDENTIFICACIÓN DEL PROBLEMA / OPORTUNIDAD

1- Presentación y fundamentación del problema / oportunidad de mejora

El comercio electrónico ha transformado radicalmente la forma en que las personas adquieren bienes y servicios. En particular, el sector de supermercados ha experimentado una rápida digitalización, impulsada por el cambio en los hábitos de consumo, el avance tecnológico y la creciente necesidad de conveniencia por parte de los usuarios. Supermercados como Lider.cl, perteneciente a Walmart Chile, han consolidado su presencia en el mercado digital ofreciendo miles de productos de consumo masivo a través de una plataforma online que opera en gran parte del territorio nacional. Sin embargo, este avance ha traído consigo una serie de desafíos estructurales, operativos y estratégicos.

Uno de los principales retos de plataformas como Lider.cl consiste en comprender y anticipar el comportamiento de compra de sus usuarios para ofrecer una experiencia personalizada, eficiente y alineada con sus necesidades reales. A pesar de contar con extensas bases de datos que almacenan el historial de compras, navegación y preferencias de miles de clientes, estos datos no siempre son procesados de manera óptima para generar conocimiento que oriente las decisiones comerciales.

Actualmente, la estrategia comercial de muchos supermercados online se basa en segmentaciones generales y promociones masivas que no necesariamente responden al perfil o comportamiento individual del cliente. Las ofertas tienden a estar dirigidas a públicos amplios, sin considerar asociaciones entre productos, hábitos de compra, preferencias por marcas, horarios habituales de compra u otras variables de segmentación relevantes. Esta falta de personalización puede

traducirse en una experiencia de usuario pobre, baja tasa de conversión, disminución del ticket promedio y, en última instancia, pérdida de lealtad del cliente. Además, la ausencia de herramientas analíticas específicas para detectar patrones de compra conlleva otros problemas operativos. Por ejemplo, si no se identifican productos que suelen comprarse juntos, se pierde la oportunidad de implementar estrategias efectivas de ventas cruzadas o de diseñar combos promocionales inteligentes. También se dificulta la planificación de stock y logística, ya que no se anticipa adecuadamente la demanda complementaria asociada a ciertos productos. Asimismo, sin una segmentación robusta de los clientes basada en sus comportamientos de compra, resulta más complejo priorizar campañas, asignar recursos de marketing o definir canales de contacto efectivos.

A nivel estratégico, esta carencia limita la capacidad de la empresa para desarrollar una cultura organizacional basada en la toma de decisiones fundamentada en datos. En mercados digitales tan dinámicos como el actual, la capacidad de una empresa para adaptarse, anticiparse y personalizar su oferta constituye una ventaja competitiva crítica. Las compañías que no integran procesos de análisis de datos avanzados corren el riesgo de quedar rezagadas frente a competidores más ágiles y orientados a lo digital.

Frente a esta situación, se detecta una oportunidad concreta de mejora: implementar un sistema de análisis basado en técnicas de minería de datos que permita identificar patrones de comportamiento de compra entre los clientes y extraer reglas de asociación entre productos. Este sistema puede ser aplicado a los datos públicos de ventas para simular su integración en la operación de Lider.cl, permitiendo así generar estrategias comerciales más precisas, fundamentadas en el comportamiento real de los consumidores.

A través del uso de algoritmos como Apriori y técnicas de clustering, se pueden obtener resultados de alto valor estratégico. Las reglas de asociación permitirán identificar productos que suelen comprarse juntos, facilitando la creación de recomendaciones automáticas o paquetes promocionales. Por otro lado, el agrupamiento de clientes en perfiles homogéneos permitirá diseñar campañas personalizadas según patrones reales de consumo, optimizar la gestión de inventario y mejorar la relación con el cliente.

En suma, la problemática identificada se relaciona con la brecha existente entre la disponibilidad de datos transaccionales y su uso efectivo para la toma de decisiones estratégicas. Esta brecha representa tanto un problema actual como una gran oportunidad de mejora. Implementar soluciones analíticas basadas en minería de datos no solo permitirá resolver los problemas mencionados, sino que también posicionará a la organización en un camino de madurez digital que puede ser ampliado a otras áreas, como recomendación de productos, predicción de demanda y personalización dinámica de interfaces.

2- Descripción de problemas / oportunidades de mejora

A partir de la fundamentación del problema expuesta anteriormente, se identifican diversos aspectos críticos dentro del funcionamiento actual de supermercados online como Lider.cl, que reflejan tanto limitaciones operativas como estratégicas. A continuación, se presentan de manera cualitativa los principales problemas y las oportunidades de mejora que emergen de su análisis:

Problemas identificados:

1. **Falta de personalización en la experiencia de compra.** La plataforma no adapta dinámicamente su oferta en función de las preferencias o patrones de

compra de cada cliente, lo que genera una experiencia de navegación genérica. Esto limita la capacidad de captar el interés del usuario y disminuir la tasa de abandono del carrito.

2. **Campañas de marketing no segmentadas.** Las promociones y ofertas son diseñadas de manera uniforme para toda la base de clientes, sin una estrategia diferenciada por tipo de consumidor. Esta falta de segmentación reduce la efectividad de las campañas y puede traducirse en un uso ineficiente del presupuesto publicitario.

3. **Pérdida de oportunidades en ventas cruzadas.** Al no identificar productos que suelen comprarse juntos, se desaprovechan instancias para generar ventas adicionales mediante recomendaciones automáticas o combos inteligentes.

4. **Gestión ineficiente del inventario y la logística**. La ausencia de información sobre patrones de compra dificulta prever con precisión la demanda asociada a productos complementarios, lo que puede conducir a sobrestock o quiebres de stock en productos clave.

5. **Limitada fidelización de clientes.** La falta de acciones personalizadas basadas en comportamiento individual disminuye las posibilidades de retener al cliente en el largo plazo. La experiencia poco adaptada puede hacer que el consumidor migre hacia plataformas más inteligentes y receptivas.

Oportunidades de mejora:

1. **Aplicación de minería de datos para identificar reglas de asociación**. A través de algoritmos como Apriori, es posible descubrir patrones de compra

y asociaciones entre productos, lo que permitiría generar recomendaciones relevantes y mejorar el ticket promedio.

2. **Segmentación de clientes basada en comportamiento de compra.** Mediante técnicas de clustering, los clientes pueden agruparse en perfiles definidos, lo cual facilitará campañas de marketing personalizadas y más efectivas.

3. **Mejoras en la planificación de promociones y combos de productos.** Conociendo los productos que se adquieren en conjunto, se pueden diseñar promociones que respondan a la lógica real de compra de los usuarios, incrementando la tasa de conversión.

4. **Optimización del inventario mediante predicción de demanda asociada**. Identificar la demanda complementaria esperada permite planificar con mayor precisión la logística y los pedidos, reduciendo pérdidas y mejorando la eficiencia.

5. **Avance en la cultura de toma de decisiones basada en datos**. La integración de herramientas analíticas no solo resuelve problemas operativos, sino que también impulsa una transformación digital en la forma en que se gestiona el negocio, fomentando el uso estratégico de la información.

3- Identificación cuantitativa de problemas (ishikawa / árbol de oportunidades)

Introducción y justificación del uso de herramientas

Para abordar de manera estructurada el problema identificado en este proyecto, se utilizaron dos herramientas de análisis ampliamente reconocidas en ingeniería: el Diagrama de Ishikawa y el Árbol de Oportunidades. Ambas permiten representar y descomponer los factores que explican el fenómeno observado, así como visualizar sus posibles efectos.

El Diagrama de Ishikawa fue aplicado para identificar las causas principales que originan la falta de análisis automatizado de patrones de compra y segmentación de clientes en la operación online de Lider.cl. Esta herramienta permite organizar las causas en categorías como personas, métodos, tecnología, información, gestión y medición, proporcionando una visión integral del problema.

Por su parte, el Árbol de Oportunidades complementa este análisis al proyectar los efectos positivos esperados si se intervienen dichas causas. A través de una estructura de raíces (causas), tronco (problema) y copa (efectos), se representa de forma clara el impacto potencial que tendría la aplicación de minería de datos en la plataforma.

Ambas herramientas fueron seleccionadas por su aplicabilidad al entorno organizacional de una plataforma de e-commerce y por su valor para orientar decisiones estratégicas a partir del análisis de datos.

Descripción de causas según Diagrama de Ishikawa

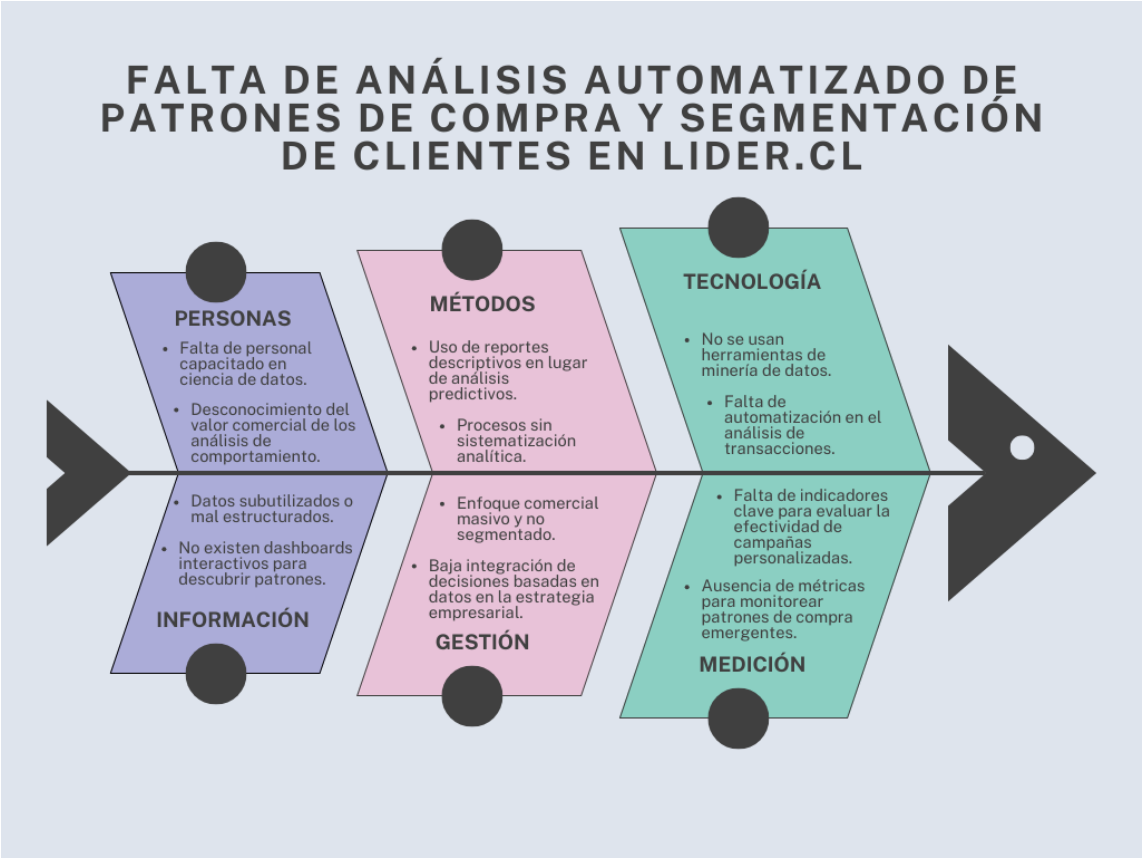

Figura 1: Diagrama de Ishikawa. Causas que explican la falta de análisis automatizado en la plataforma online de Lider.cl.

A continuación, se detallan las principales causas del problema identificado, clasificadas en seis categorías analizadas mediante el Diagrama de Ishikawa:

1. Personas:

- Falta de personal especializado en ciencia de datos: La empresa no cuenta con perfiles técnicos capacitados para implementar modelos de análisis avanzado.
- Desconocimiento de herramientas de minería de datos: Muchos actores relevantes en la gestión comercial desconocen el potencial de estas técnicas.

2. Métodos:

- Enfoque tradicional en reportes descriptivos: El análisis de ventas se basa en informes básicos sin segmentación avanzada ni modelado de comportamiento.
- Ausencia de procesos analíticos formalizados: No existen flujos estandarizados para aprovechar los datos transaccionales en la toma de decisiones.

3. Tecnología:

- Falta de herramientas específicas para minería de datos: La plataforma no integra soluciones como Apriori, clustering u otras técnicas similares.
- Baja automatización del análisis: El procesamiento de información se realiza manualmente o con herramientas no especializadas.

4. Información:

- Subutilización de los datos históricos: Aunque existen registros de compras, estos no se explotan para extraer conocimiento estratégico.
- Datos no organizados para análisis: La estructura actual de almacenamiento dificulta el uso eficiente de la información.

5. Gestión:

- Toma de decisiones basada en intuición o experiencia: Las campañas de marketing y promociones no se basan en datos concretos de comportamiento.
- Foco en tareas operativas por sobre el análisis estratégico: Se priorizan actividades de corto plazo sobre iniciativas analíticas de largo plazo.

6. Medición:

- Falta de indicadores para evaluar patrones de compra: No se han definido métricas que permitan monitorear la efectividad de acciones comerciales basadas en comportamiento.
- Ausencia de retroalimentación analítica: No se mide el impacto de las decisiones relacionadas con segmentación y recomendaciones.

Análisis del Árbol de Oportunidades

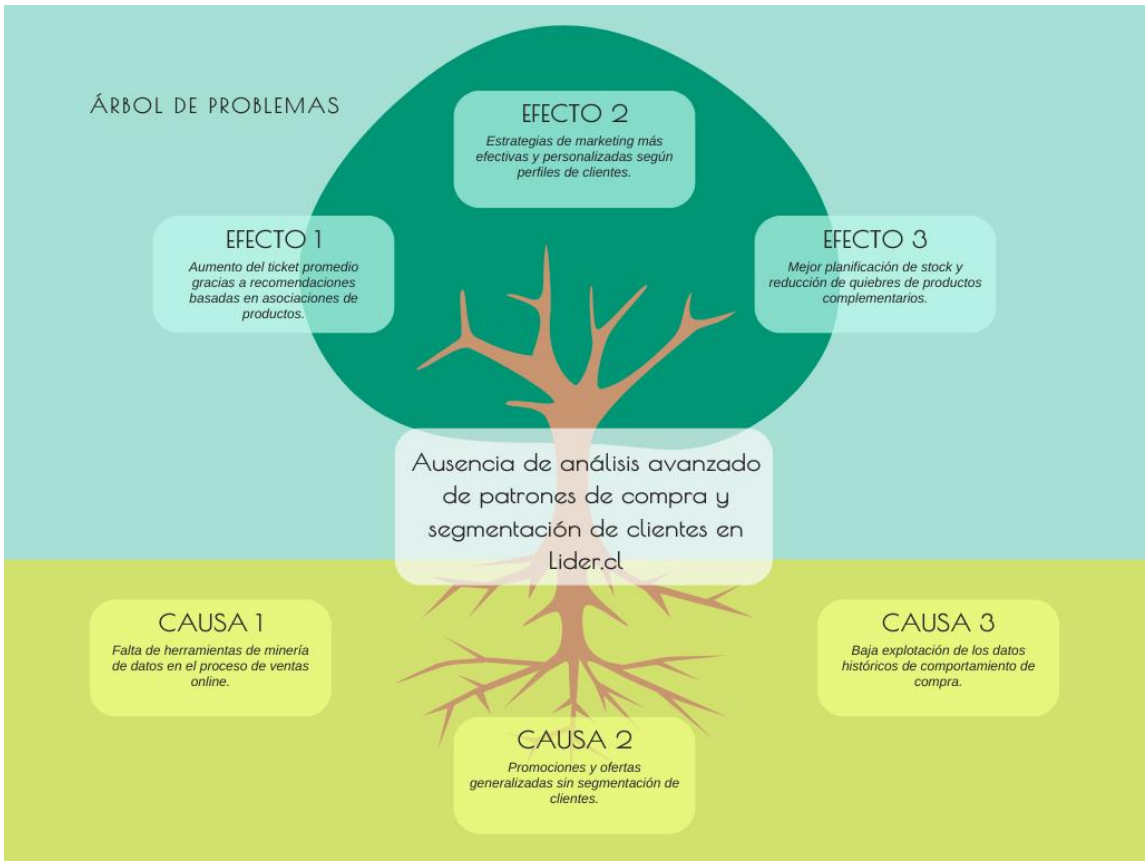

Figura 2: Árbol de Oportunidades. Visualización del problema, sus causas y los efectos positivos que se esperan al aplicar técnicas de minería de datos en Lider.cl.

El Árbol de Oportunidades permite visualizar los beneficios esperados al intervenir las causas identificadas que limitan el análisis avanzado de datos en la operación

de Lider.cl. Esta herramienta se construye a partir del problema central ya definido, sus causas principales (raíces) y los efectos positivos proyectados (copa del árbol).

Entre las causas clave se identifican la falta de herramientas de análisis, la subutilización de los datos históricos y la aplicación de estrategias de marketing generalizadas. Estas limitaciones afectan la capacidad de la empresa para adaptar su oferta a los hábitos de consumo de los clientes y responder con agilidad a la demanda real del mercado.

Al aplicar técnicas de minería de datos —como reglas de asociación para descubrir productos que se compran juntos y clustering para segmentar perfiles de clientes— se espera lograr efectos concretos. Entre ellos destacan: un aumento del ticket promedio a través de ventas cruzadas más efectivas, una mejora en la personalización de promociones, una mayor fidelización de clientes, y una planificación de inventario más precisa basada en patrones de compra reales.

Este modelo proyecta, por tanto, una transformación estratégica de la operación comercial, al pasar de un enfoque reactivo y masivo a uno proactivo y personalizado, guiado por datos. El Árbol de Oportunidades no solo justifica la solución propuesta, sino que también articula claramente el valor agregado que puede generar este proyecto en términos de eficiencia, satisfacción del cliente y ventaja competitiva.

Estudios recientes en plataformas de retail digital demuestran que la implementación de análisis de datos segmentados y personalizados puede aumentar el ticket promedio entre un 15% y un 30%, y mejorar la tasa de retención de clientes hasta en un 25% (Song & Kim, 2022). Por tanto, no abordar las causas estructurales identificadas implicaría mantener una brecha competitiva importante frente a empresas que ya integran minería de datos en sus procesos estratégicos.

4- Objetivo general

Desarrollar un sistema de análisis de datos basado en técnicas de minería de datos que permita identificar patrones de compra y segmentar clientes en el contexto del supermercado online Lider.cl, con el fin de optimizar estrategias de venta, personalizar la experiencia del usuario y mejorar la toma de decisiones comerciales.

5- Objetivos específicos y métricas

Objetivo específico 1:

Aplicar técnicas de reglas de asociación para identificar productos que se compran frecuentemente juntos en la plataforma Lider.cl.

- Métrica: Número de reglas relevantes descubiertas con soporte y confianza superiores al umbral.
- Criterio de éxito: Al menos 10 reglas con soporte ≥ 5% y confianza ≥ 70%.

Objetivo específico 2:

Segmentar a los clientes en grupos con patrones de compra similares mediante algoritmos de clustering.

- Métrica: Cohesión de los grupos (medida con Silhouette Score).
- Criterio de éxito: Al menos 3 grupos significativos con Silhouette Score ≥ 0.5.

Objetivo específico 3:

Interpretar los resultados obtenidos para generar recomendaciones aplicables a la estrategia comercial de Lider.cl.

- Métrica: Número de recomendaciones estratégicas generadas.
- Criterio de éxito: Al menos 5 recomendaciones viables alineadas con las asociaciones y segmentos detectados.

6- Métricas de los objetivos específicos

Objetivo específico	Métrica	Criterio de éxito
Aplicar técnicas de reglas de asociación para identificar productos que se compran frecuentemente juntos.	Número de reglas con soporte y confianza altos.	≥ 10 reglas con soporte ≥ 5% y confianza ≥ 70%.
Segmentar a los clientes según patrones de compra mediante clustering.	Cohesión de grupos (Silhouette Score).	≥ 3 grupos con Silhouette Score ≥ 0.5.
Interpretar los resultados y generar recomendaciones comerciales.	Cantidad de recomendaciones viables.	≥ 5 recomendaciones estratégicas.

Tabla 1: Métricas de los objetivos específicos.

7- Limitaciones y alcances del proyecto

Alcances

El presente proyecto se enfoca en el análisis de datos transaccionales provenientes de un conjunto de datos públicos representativo del comportamiento de compra en supermercados online, simulando su aplicación en la operación de **Lider.cl**. El análisis contempla la aplicación de técnicas de minería de datos específicas, como **reglas de asociación** y **clustering**, con el objetivo de generar conocimiento útil para la toma de decisiones comerciales.

Las actividades del proyecto incluyen:

- Limpieza y preparación del conjunto de datos.
- Aplicación del algoritmo A priori para identificar asociaciones entre productos.

- Implementación de algoritmos de agrupamiento de clientes.
- Interpretación de resultados y generación de recomendaciones accionables.
- Elaboración de un informe técnico con conclusiones y oportunidades de aplicación.

Limitaciones

- **Datos simulados**: No se utilizarán datos reales de Lider.cl por restricciones de acceso, sino datasets públicos que simulan una operación similar. Esto limita el nivel de personalización directa de los hallazgos.

- **Enfoque exploratorio**: El objetivo del proyecto es realizar un análisis exploratorio y no construir un sistema productivo completo. No se implementarán motores de recomendación en producción ni sistemas de automatización real.

- **Restricciones de tiempo y recursos**: Debido al carácter académico del portafolio, el proyecto se desarrollará en un marco temporal limitado y sin presupuesto destinado a infraestructura de análisis avanzado o herramientas comerciales.

- **Generalización de resultados**: Las recomendaciones generadas a partir de los análisis estarán acotadas al comportamiento observado en el conjunto de datos analizado y no necesariamente serán extrapolables a otras plataformas sin ajustes contextuales.

8- Normativa y leyes asociadas al proyecto (si aplica)

Dado que este proyecto trabaja con datos públicos y anónimos, no se vulnera ninguna normativa de protección de datos personales ni privacidad del consumidor. No obstante, en un escenario real de implementación con datos privados de clientes, se deberían considerar los siguientes marcos normativos:

- Ley N.º 19.628 sobre Protección de la Vida Privada (Chile): Regula el tratamiento de datos personales en el país, estableciendo que toda recolección, almacenamiento o análisis de datos debe contar con el consentimiento del titular.

- Políticas internas de privacidad de Lider.cl: En un entorno real, cualquier análisis de comportamiento de clientes debe respetar los términos y condiciones de uso de datos establecidos por la empresa y comunicados al usuario.

El presente proyecto, al utilizar datos simulados o públicos, no incurre en infracciones legales y se mantiene dentro de un marco ético y académico.

CAPÍTULO III: MARCO METODOLÓGICO

1- Metodología

Este proyecto considera dos niveles metodológicos complementarios: por una parte, una metodología de desarrollo específica para la aplicación de técnicas de minería de datos; y por otra, un enfoque de gestión de proyecto que permita planificar y controlar su ejecución de manera estructurada y flexible.

Metodología de desarrollo

Para el desarrollo del análisis de datos se adoptará la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*), ampliamente utilizada en proyectos de ciencia de datos por su enfoque estructurado y adaptable. Esta metodología se compone de seis etapas que guían el proceso desde la definición del problema hasta la implementación de soluciones basadas en datos:

1. Comprensión del negocio: Se define el problema central asociado a la falta de análisis estructurado de datos transaccionales en supermercados online como Lider.cl, y se establecen los objetivos analíticos.

2. Comprensión de los datos: Se selecciona un conjunto de datos representativo del comportamiento de compra en entornos de e-commerce. Se realiza una evaluación inicial de su estructura, calidad y potencial analítico.

3. Preparación de los datos: Se lleva a cabo la limpieza, transformación y codificación de los datos para dejarlos en condiciones óptimas para su análisis.

4. Modelado: Se implementan técnicas de minería de datos, en este caso reglas de asociación mediante el algoritmo Apriori y segmentación mediante K-means.

5. Evaluación: Se validan los resultados obtenidos a partir de métricas específicas, como soporte y confianza para reglas de asociación, y Silhouette Score para evaluar la cohesión de los grupos de clientes.

6. Despliegue (simulado): Se presentan los hallazgos en un informe técnico y se diseña un prototipo visual que representa cómo podrían utilizarse los resultados en una plataforma de supermercado online.

Este enfoque permite abordar el análisis de manera rigurosa, iterativa y orientada a resultados aplicables.

Metodología de gestión

En cuanto a la gestión del proyecto, se opta por una metodología híbrida que combina elementos del enfoque ágil Scrum con componentes del marco PMBOK. Esta elección responde a la necesidad de mantener una planificación clara con entregables definidos, al mismo tiempo que se permite flexibilidad para adaptar el desarrollo según los resultados obtenidos en cada etapa.

Scrum se utilizará para dividir el trabajo en iteraciones semanales, priorizando tareas como la exploración inicial de los datos, el desarrollo de los modelos analíticos y la interpretación de los resultados. Por su parte, los principios de PMBOK se aplicarán para estructurar el alcance, los riesgos, la calidad y la documentación del proyecto.

Esta combinación metodológica resulta especialmente adecuada para proyectos de análisis de datos en entornos académicos, ya que permite un equilibrio entre planificación y adaptabilidad.

Este conjunto de herramientas ha sido validado y reforzado durante la participación en el curso de Minería de Datos de la plataforma IBM Skills Network, en el cual se aplicaron técnicas similares de análisis sobre conjuntos de datos reales utilizando Python, CRISP-DM y herramientas como Jupyter Notebook. Esta experiencia previa permite abordar el desarrollo del presente proyecto con un enfoque práctico y metodológicamente fundamentado.

2- Herramientas y ambiente de desarrollo

El desarrollo del proyecto se realizará en un entorno técnico orientado al análisis de datos y la minería de patrones de comportamiento. Las herramientas y tecnologías seleccionadas han sido definidas en función de su capacidad para soportar procesos de preparación, modelado, evaluación e interpretación de datos de manera eficiente y reproducible.

Lenguajes de programación

- Python: Será el lenguaje principal de implementación. Su elección se justifica por su amplia comunidad, flexibilidad y por la disponibilidad de librerías especializadas para minería de datos como pandas, mlxtend, scikit-learn y matplotlib. Python permite realizar tanto la preparación de datos como la aplicación de modelos y la visualización de resultados de forma integrada.

- R: Se utilizará como complemento para análisis estadísticos y representación gráfica avanzada. R destaca por su precisión en cálculos estadísticos y por

librerías como ggplot2 y dplyr, que permiten generar visualizaciones de alta calidad para la interpretación de resultados.

Entorno de desarrollo

- Google Colab: Será el entorno principal de trabajo para Python. Permite la ejecución de notebooks basados en Jupyter directamente desde la nube, sin necesidad de instalación local, y ofrece compatibilidad con bibliotecas especializadas de análisis de datos. Además, facilita el trabajo incremental y la documentación del proceso analítico en un solo archivo.

- RStudio: Se empleará para los análisis realizados en R. Es un entorno especializado para dicho lenguaje, que permite organizar scripts, gráficos y documentación de forma estructurada.

- Visual Studio Code (VSC): Se utilizará como editor complementario para la organización del proyecto, control de versiones y edición de scripts en Python. Su flexibilidad y capacidad de integración con Git lo convierten en una herramienta eficiente para el desarrollo modular del código.

Repositorio y control de versiones

- GitHub: Se utilizará como repositorio para almacenar versiones del código, gráficos y reportes generados durante el desarrollo del proyecto. Su uso garantiza trazabilidad y facilita el trabajo incremental.

Este conjunto de herramientas ha sido validado y reforzado durante la participación en el curso de Minería de Datos de la plataforma IBM Skills Network, en el cual se aplicaron técnicas similares de análisis sobre conjuntos de datos reales utilizando

Python, CRISP-DM, Jupyter Notebook y control de versiones con GitHub. Esta experiencia previa permite abordar el desarrollo del presente proyecto con un enfoque práctico y metodológicamente fundamentado.

3- Gestión de proyecto

La gestión del proyecto se organizará en torno a un enfoque híbrido, integrando elementos de metodologías ágiles (Scrum) con principios estructurados del marco PMBOK, con el fin de garantizar tanto la flexibilidad del desarrollo como el cumplimiento de entregables formales. Esta estrategia dual permite abordar un proyecto de carácter analítico dentro de un contexto académico de duración limitada, facilitando el control y la adaptabilidad durante su ejecución.

Estructura general de gestión

Se contempla una planificación inicial basada en un cronograma general de etapas, complementado con iteraciones semanales (sprints) que permitirán abordar tareas específicas de forma incremental. Las principales fases del proyecto se estructuran de la siguiente manera:

1. Definición del problema y objetivos: Análisis del contexto, formulación del problema y elaboración de los objetivos generales y específicos del proyecto.

2. Levantamiento y comprensión de los datos: Selección del dataset, revisión de sus atributos, comprensión de su estructura y evaluación inicial de calidad.

3. Preparación de datos y desarrollo exploratorio: Limpieza, transformación y codificación de datos, acompañada de visualizaciones iniciales y análisis descriptivo.

4. Modelado y aplicación de técnicas de minería de datos: Implementación de reglas de asociación y segmentación de clientes mediante clustering.

5. Evaluación de resultados y redacción de hallazgos: Análisis de los resultados obtenidos, interpretación y validación de patrones, redacción de conclusiones y elaboración del informe final.

6. Desarrollo de prototipo y visualización de resultados: Diseño de representaciones visuales de los hallazgos que podrían incorporarse en plataformas de supermercado online.

7. Cierre y documentación: Consolidación de resultados, revisión final, carga en repositorio y entrega del portafolio completo.

Responsabilidades y seguimiento

A lo largo del proyecto se realizarán sesiones semanales de revisión del avance, con énfasis en:

- Validación del cumplimiento de objetivos por fase.
- Resolución de obstáculos técnicos o metodológicos.
- Documentación progresiva de código, resultados y decisiones tomadas.

Adicionalmente, se mantendrá el control de versiones del código fuente a través de GitHub, lo que permitirá asegurar la trazabilidad del trabajo desarrollado.

4- Plan de gestión (riesgos, calidad y testing)

Durante el desarrollo del proyecto se identifican algunos riesgos críticos que podrían comprometer su avance o la validez de los resultados. A continuación, se presenta una tabla con los principales riesgos, su nivel de impacto y las estrategias de mitigación consideradas:

Riesgo identificado	Impacto	Probabilidad	Estrategia de mitigación
Baja calidad del dataset seleccionado	Alto	Media	Evaluar múltiples datasets y aplicar técnicas de limpieza y validación estadística.
Dificultades en la implementación de algoritmos	Medio	Alta	Consultar documentación técnica, repositorios académicos y buscar apoyo en foros especializados.
Exceso de carga académica externa al proyecto	Alto	Alta	Definir cronograma realista y reservar bloques fijos de tiempo semanal para avanzar.
Resultados no interpretables o poco relevantes	Medio	Media	Ajustar los parámetros de los modelos y, si es necesario, redefinir el enfoque exploratorio.

Tabla 2: Tabla con los principales riesgos, su nivel de impacto y las estrategias de mitigación consideradas.

Este plan permite anticipar escenarios adversos y asegurar la continuidad del trabajo en condiciones cambiantes.

Gestión de la calidad

La calidad del proyecto se abordará desde dos dimensiones:

- **Calidad del proceso analítico**: Se garantizará mediante la trazabilidad del flujo de trabajo (limpieza, modelado, validación), el uso de herramientas estandarizadas (Python, R, librerías científicas) y la documentación clara del código y decisiones tomadas.

- **Calidad del producto entregable**: Se controlará a través de revisiones periódicas del informe escrito, revisión de formato, cumplimiento de objetivos definidos y validación de que los resultados sean coherentes con el planteamiento inicial del problema.

Además, se establecerán puntos de control semanales para revisar avances, detectar errores metodológicos tempranos y mantener la coherencia técnica y argumentativa del portafolio.

Plan de pruebas (testing)

Para validar el funcionamiento correcto de los modelos aplicados se consideran las siguientes pruebas:

- **Para reglas de asociación (Apriori)**: Revisión del soporte y confianza de las reglas generadas, validación cruzada de los pares frecuentes más relevantes y exclusión de resultados triviales o no útiles desde el punto de vista comercial.

- **Para clustering (K-means)**: Evaluación de la calidad del agrupamiento mediante la métrica de Silhouette Score, análisis visual de la separación de los clústeres y revisión de la coherencia de los grupos identificados respecto a las variables analizadas.

En ambos casos, se aplicará el principio de interpretación cualitativa de los resultados, verificando que los patrones descubiertos puedan transformarse en conocimiento útil para la toma de decisiones comerciales.

5- Ingeniería de requerimientos (funcionales y no funciones)

Para estructurar el desarrollo del proyecto y anticipar los aspectos clave que debe cumplir la solución analítica propuesta, se identifican y clasifican los requerimientos del sistema en dos categorías: funcionales y no funcionales. Esta definición permite establecer expectativas claras respecto a las capacidades que debe ofrecer el análisis, los resultados esperados y las condiciones bajo las cuales estos serán útiles y viables.

Requerimientos funcionales

Los requerimientos funcionales se refieren a las funciones concretas que el sistema o análisis debe ser capaz de realizar. En el contexto de este proyecto, se identifican los siguientes:

- RF1: El sistema debe permitir cargar, procesar y transformar un conjunto de datos de compras en supermercados online.
- RF2: El sistema debe generar reglas de asociación entre productos utilizando el algoritmo Apriori.

- RF3: El sistema debe identificar grupos de clientes mediante clustering con el algoritmo K-means.
- RF4: El sistema debe visualizar de forma comprensible los patrones obtenidos, ya sea en tablas o gráficos.
- RF5: El sistema debe permitir interpretar los resultados obtenidos de forma que se puedan extraer conclusiones accionables para el negocio.

Requerimientos no funcionales

Los requerimientos no funcionales definen atributos de calidad que debe cumplir la solución desarrollada. Para este proyecto, se consideran los siguientes:

- RNF1: El sistema debe estar implementado con herramientas abiertas y gratuitas (Python, R, Colab, etc.).
- RNF2: El código debe ser modular, reutilizable y documentado.
- RNF3: El tiempo de ejecución de los algoritmos debe ser razonable y no superar los 2 minutos en un entorno estándar.
- RNF4: La visualización de los resultados debe ser clara y comprensible para un público no técnico.
- RNF5: Los resultados obtenidos deben ser trazables y reproducibles en cualquier entorno compatible.

Esta especificación de requerimientos permite guiar el desarrollo de la solución y validar posteriormente si los objetivos del proyecto han sido cumplidos tanto en términos funcionales como de calidad operativa.

6- PMBOK y planes de proyecto

Siguiendo las buenas prácticas propuestas por el estándar PMBOK, se han definido distintos planes de gestión que permiten estructurar el desarrollo del proyecto desde una perspectiva integral. Estos planes contribuyen a reducir la incertidumbre, alinear expectativas y facilitar el cumplimiento de los objetivos en tiempo y forma.

Plan de alcance

El alcance del proyecto se limita al desarrollo de un sistema analítico que permita identificar patrones de compra y segmentar clientes a partir de datos transaccionales representativos de un supermercado online. No se contempla la integración directa del sistema en una plataforma comercial real ni el acceso a bases de datos privadas. El resultado final será un análisis replicable, documentado y visualmente representado, acompañado de un informe técnico y un prototipo conceptual.

Este alcance garantiza la viabilidad del proyecto dentro de los plazos establecidos y mantiene el enfoque en el desarrollo de capacidades analíticas a partir de datos públicos o simulados.

Plan de riesgos

Los riesgos críticos ya identificados en la sección anterior se integran formalmente en el plan de gestión. Cada riesgo cuenta con una estrategia preventiva y/o correctiva, y se revisará periódicamente durante los puntos de control establecidos en el cronograma.

Se mantendrá un registro simple pero efectivo de los eventos de riesgo, su impacto y las acciones adoptadas, utilizando una tabla de seguimiento integrada al repositorio del proyecto.

Plan de calidad

La calidad del proyecto será gestionada mediante dos mecanismos principales:

1. Validación cruzada de resultados técnicos (reglas, clústeres, visualizaciones) con respecto a los objetivos definidos.
2. Revisión formal del informe escrito y del código, asegurando que cumplan con criterios de claridad, coherencia, reproducibilidad y trazabilidad.

El cumplimiento de los requerimientos funcionales y no funcionales definidos previamente servirá como base para evaluar la calidad del entregable final.

Plan de comunicaciones

Dado el carácter académico del proyecto, las comunicaciones se organizan en dos niveles:

- A nivel individual, se establecerán revisiones semanales del avance, registro de decisiones y actualizaciones internas en el repositorio.
- A nivel institucional, se considerarán los canales formales definidos por la asignatura para la entrega de avances, retroalimentación y evaluaciones.

El plan de comunicaciones asegura que los actores involucrados tengan acceso oportuno a la información clave del proyecto.

7- Diseño y arquitectura de alto nivel

El diseño del sistema se estructura considerando tanto las funcionalidades requeridas como la naturaleza del análisis basado en datos. Para ello, se emplean herramientas de modelado que permiten representar de manera clara las interacciones funcionales y los componentes principales involucrados en la solución.

Diagrama de casos de uso: El sistema desarrollado presenta una estructura simple centrada en un único usuario (analista), que ejecuta cada etapa del proceso según el modelo CRISP-DM. A continuación, se presenta una descripción de los principales casos de uso:

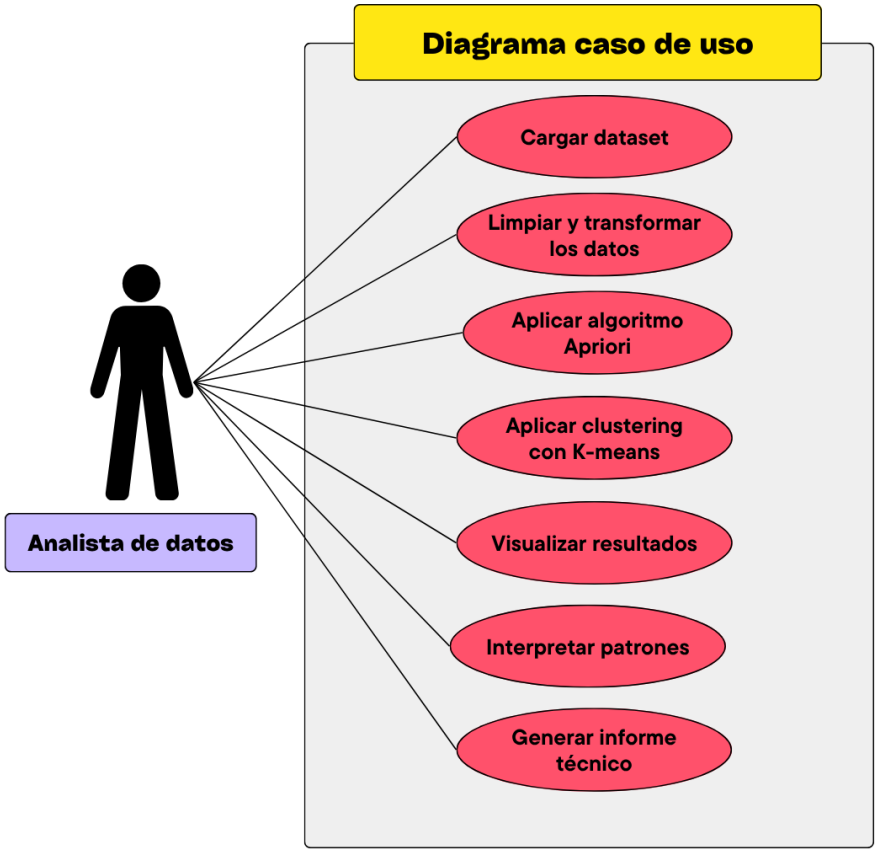

Figura 3: Diagrama de caso de uso.

Este diagrama describe la interacción entre el usuario y los componentes principales del sistema analítico.

Vistas del sistema

A continuación, se describen las vistas requeridas para representar la arquitectura general de la solución, considerando su carácter analítico y académico.

Vista lógica. Representa los módulos funcionales que componen el sistema, organizados en etapas:

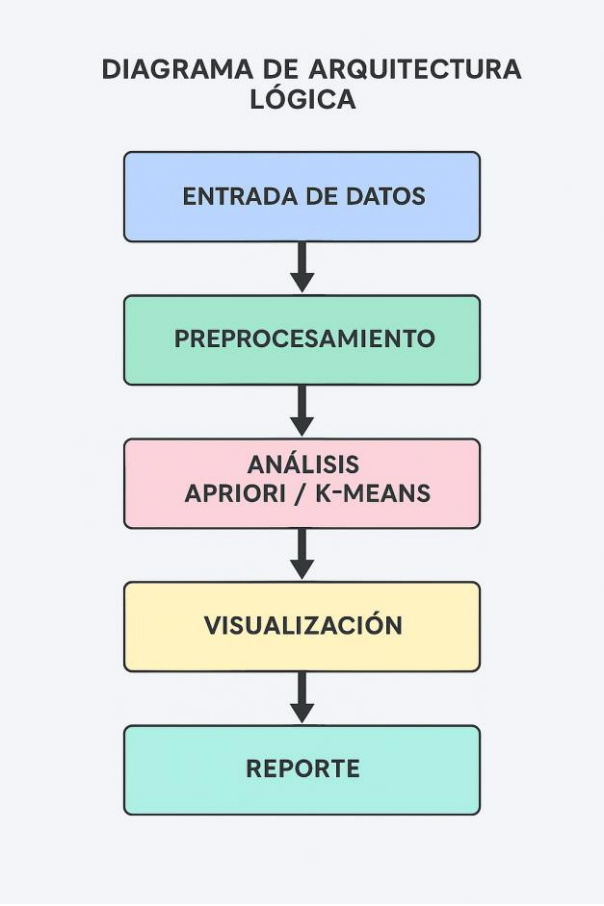

Figura 4: Diagrama de arquitectura lógica.

Vista de desarrollo. Describe cómo se organizará el código y los archivos del proyecto:

- Scripts separados por función (preprocessing.py, association.py, clustering.py).
- Notebooks para documentación interactiva (analisis_final.ipynb).
- Carpeta de recursos con visualizaciones y archivos auxiliares.
- Archivo README con instrucciones de uso.

Vista de procesos. Describe la secuencia general de ejecución:

1. Carga del conjunto de datos.
2. Limpieza y transformación.
3. Aplicación de algoritmos.
4. Generación de salidas (gráficos, tablas).
5. Análisis e interpretación de resultados.

Vista física. Describe el entorno en el cual se ejecutará la solución:

- Dispositivo local con acceso a Google Colab (Python 3).
- Navegador web con conexión a internet.
- Repositorio en GitHub para control de versiones y respaldo del proyecto.

8- Cronograma del proyecto

El cronograma del proyecto se ha diseñado considerando las fases principales definidas en la metodología CRISP-DM, combinadas con la estructura iterativa de gestión ágil basada en sprints semanales. El objetivo de esta planificación es asegurar que cada etapa del proceso analítico sea abordada de forma progresiva, controlada y con tiempos razonables.

El proyecto contempla un desarrollo total de 8 semanas, considerando la disponibilidad académica y las fechas de entrega definidas por la asignatura. Cada sprint tiene una duración estimada de una semana, permitiendo el desarrollo incremental de las tareas. A continuación, se describe el plan general del proyecto:

Semana	Fase del proyecto	Actividades clave
1	**Definición del problema**	Redacción inicial, objetivos, revisión bibliográfica
2	**Comprensión de datos**	Selección de dataset, análisis exploratorio inicial
3	**Preparación de datos**	Limpieza, codificación, tratamiento de valores nulos
4	**Modelado – Asociación**	Aplicación de Apriori, validación de reglas, soporte/confianza
5	**Modelado – Clustering**	Aplicación de K-means, evaluación con Silhouette, visualización
6	**Evaluación de resultados**	Interpretación de patrones, análisis de impacto comercial
7	**Prototipo y visualización**	Elaboración de gráficas, bosquejo de sistema conceptual
8	**Cierre y entrega**	Revisión final, correcciones, envío del informe y anexos

Tabla 3: Cronograma del proyecto.

Metodología de desarrollo y control

Cada semana será estructurada como un sprint con objetivos concretos, tareas planificadas y revisión de avance al cierre. Se utilizará una checklist semanal, además del repositorio de GitHub, para registrar avances y decisiones. Las herramientas Colab, VSC y Google Drive facilitarán la documentación colaborativa y el respaldo del proyecto.

Este cronograma garantiza un avance sostenido, con tiempos destinados a evaluación y corrección que aseguran tanto el cumplimiento del alcance como la calidad del entregable final.

9- Prototipo

Dado que el presente proyecto se enmarca en un entorno académico y se centra en el análisis de datos y generación de conocimiento, el prototipo se plantea como una representación conceptual y funcional de los resultados obtenidos, más que como un producto final desplegado en ambiente productivo.

El prototipo tendrá como objetivos principales:

- Visualizar gráficamente los patrones de compra detectados mediante el algoritmo Apriori.
- Representar los grupos de clientes identificados por el clustering (K-means) mediante gráficos de dispersión.
- Simular, a modo de ejemplo, cómo estos resultados podrían incorporarse en una plataforma online tipo supermercado.

Características del prototipo

- Se desarrollará en Google Colab, utilizando librerías como matplotlib, seaborn y plotly para generar las visualizaciones interactivas.
- Las visualizaciones incluirán:

    - Reglas de asociación frecuentes en tablas ordenadas por soporte y confianza.
    - Clústeres de clientes representados gráficamente por variables clave (como monto de compra y frecuencia).
    - Tablas resumen con insights clave por grupo de clientes.

- Se incluirá una sección final que simula cómo estas visualizaciones podrían integrarse en un dashboard o sistema de apoyo a la toma de decisiones para el área comercial o de marketing.

Finalidad

Este prototipo no será un sistema funcional conectado a Lider.cl, sino una demostración técnica y visual del valor de aplicar minería de datos en este contexto. Su propósito es comunicar de forma clara, tanto a públicos técnicos como no técnicos, el impacto potencial del análisis desarrollado en este proyecto.

CAPÍTULO IV: HALLAZGOS PRELIMINARES

Este capítulo sintetiza los hallazgos obtenidos a partir de las técnicas de *Minería de Datos* aplicadas al conjunto de datos de compras. Se utilizaron dos enfoques complementarios: reglas de asociación (Apriori) y segmentación de clientes mediante clustering (K-Means). Cada técnica permitió extraer patrones relevantes desde perspectivas distintas, entregando una visión integral del comportamiento de los usuarios.

1- Reglas de Asociación (A priori)

Las reglas de asociación permitieron descubrir relaciones frecuentes entre productos, entregando una base para estrategias como ventas cruzadas o recomendaciones personalizadas.

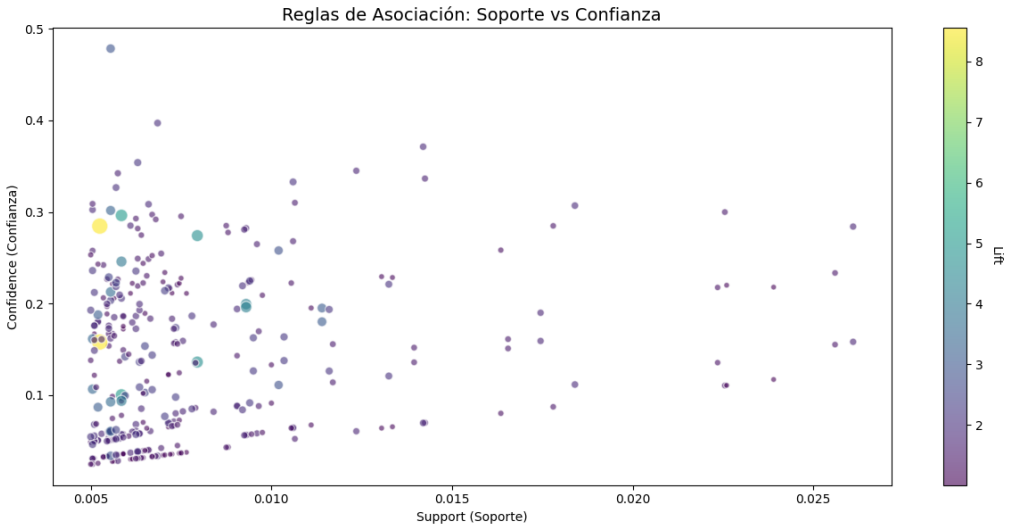

Figura 5. Reglas de asociación. Soporte vs confianza

La mayoría de las reglas tienen soporte bajo, pero confianza intermedia (alrededor de 0.2 a 0.3). Esto indica que, aunque no son compras masivas, muchas ocurren con frecuencia condicional una vez iniciado un patrón. Además, algunos puntos se

destacan con valores altos de *lift*, revelando asociaciones mucho más fuertes que el azar.

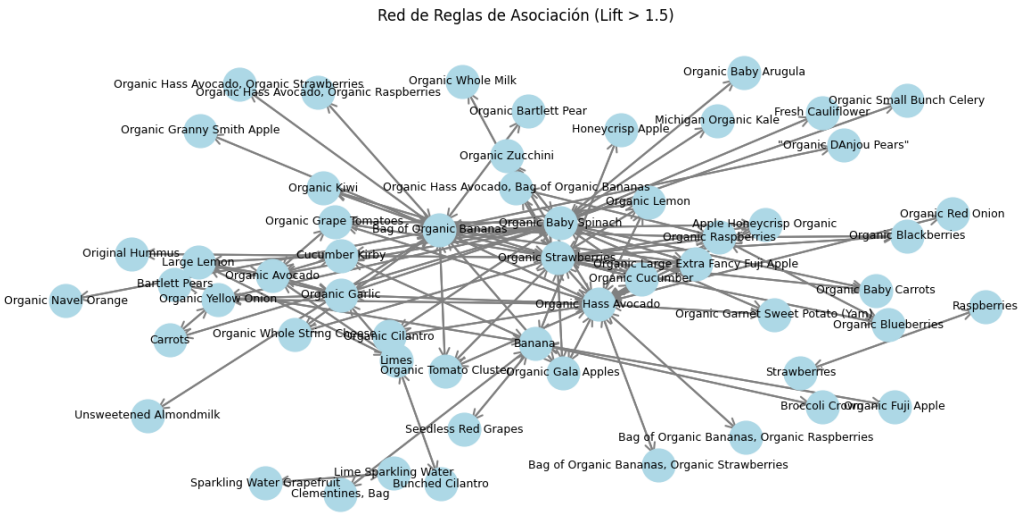

Figura 6. Red de reglas de asociación

Se observa una red de productos altamente conectados. Algunos productos como *“Organic Strawberries”*, *“Bananas”* y *“Organic Avocado”* actúan como nodos centrales. Este hallazgo es clave para definir productos gancho o de entrada en campañas de marketing o layout de tiendas.

2- Segmentación de Clientes (Clustering)

Se identificaron cinco clústeres principales de clientes según su comportamiento de compra (horario, frecuencia y tipo de productos). Cada uno representa un patrón específico de consumo.

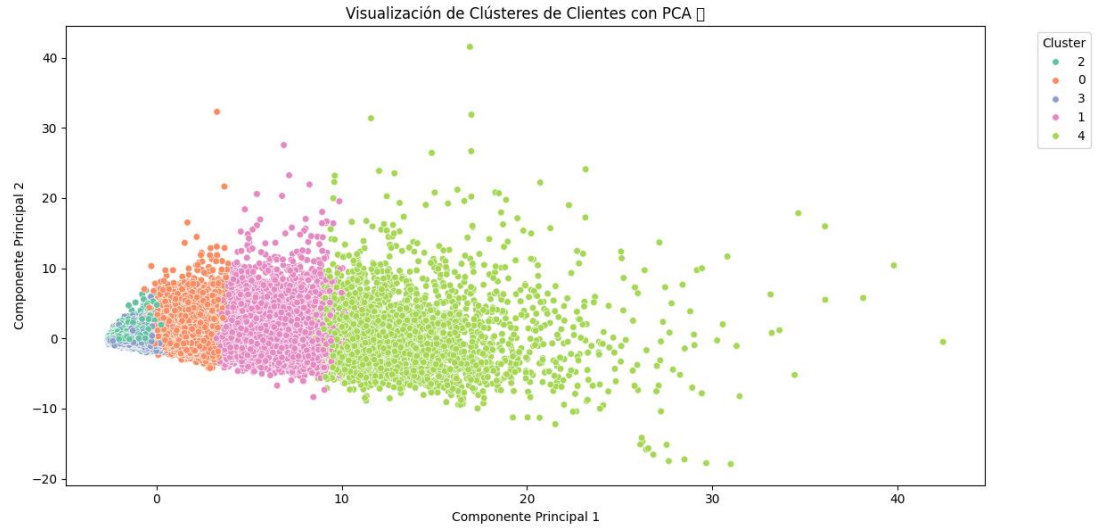

Figura 7. Visualización PCA y Distribución

El PCA muestra una separación clara entre grupos, y se confirma con una distribución desigual: los clústeres 2 y 3 concentran la mayoría de usuarios. Esto indica segmentos dominantes con patrones comunes.

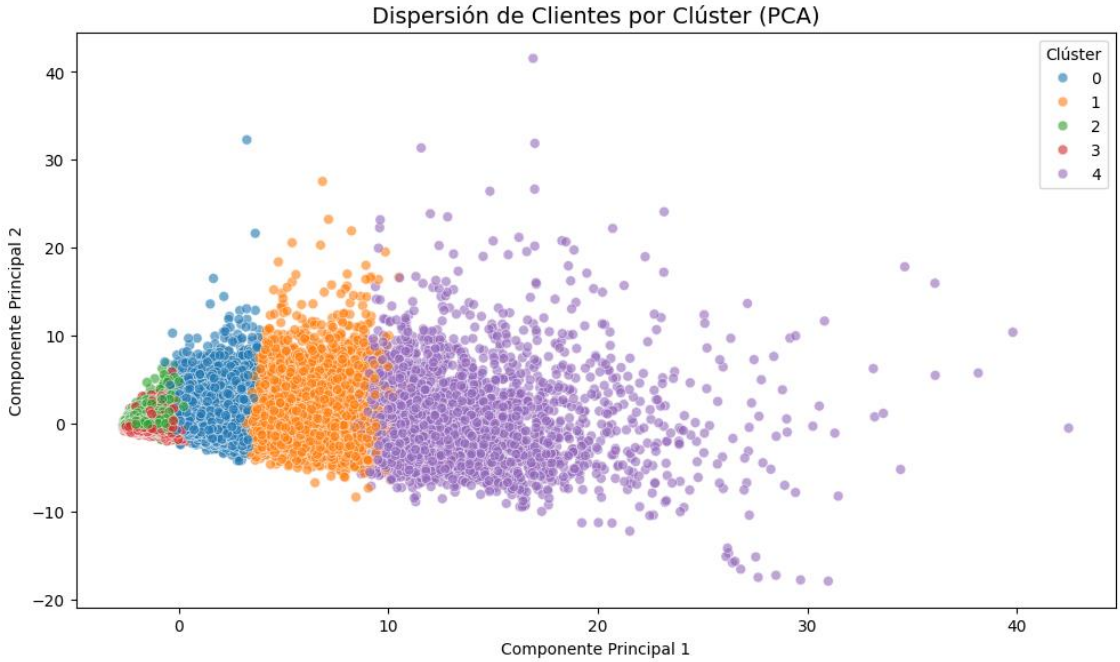

Figura 8. Dispersión de clientes por clúster (PCA)

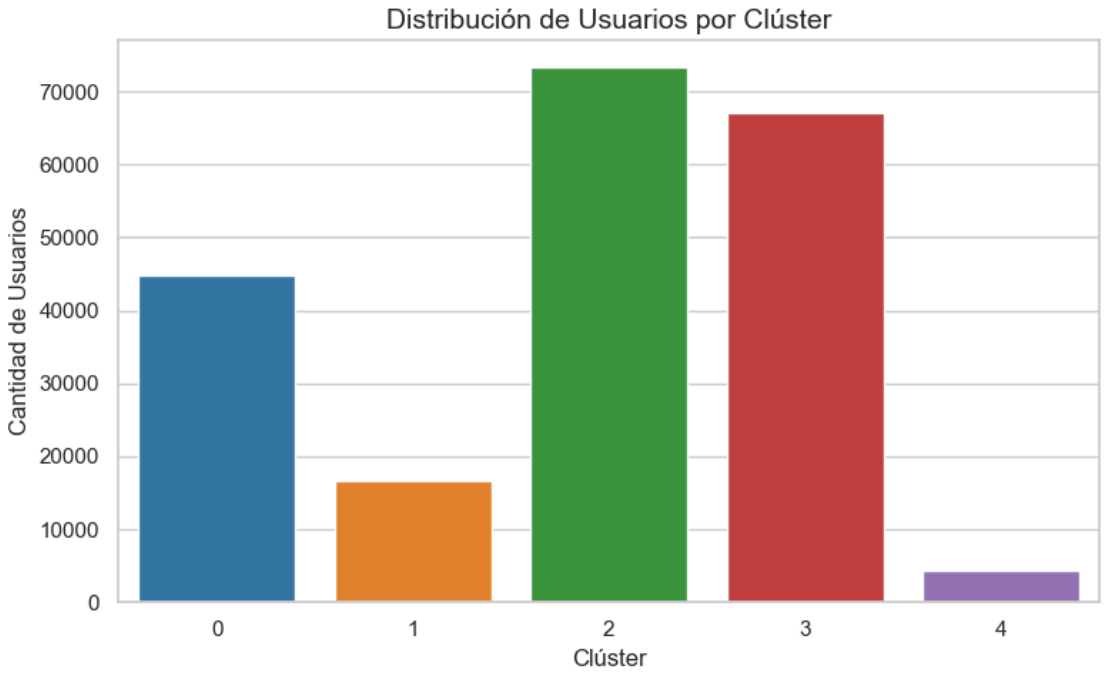

Figura 9. Distribución de usuarios por clúster

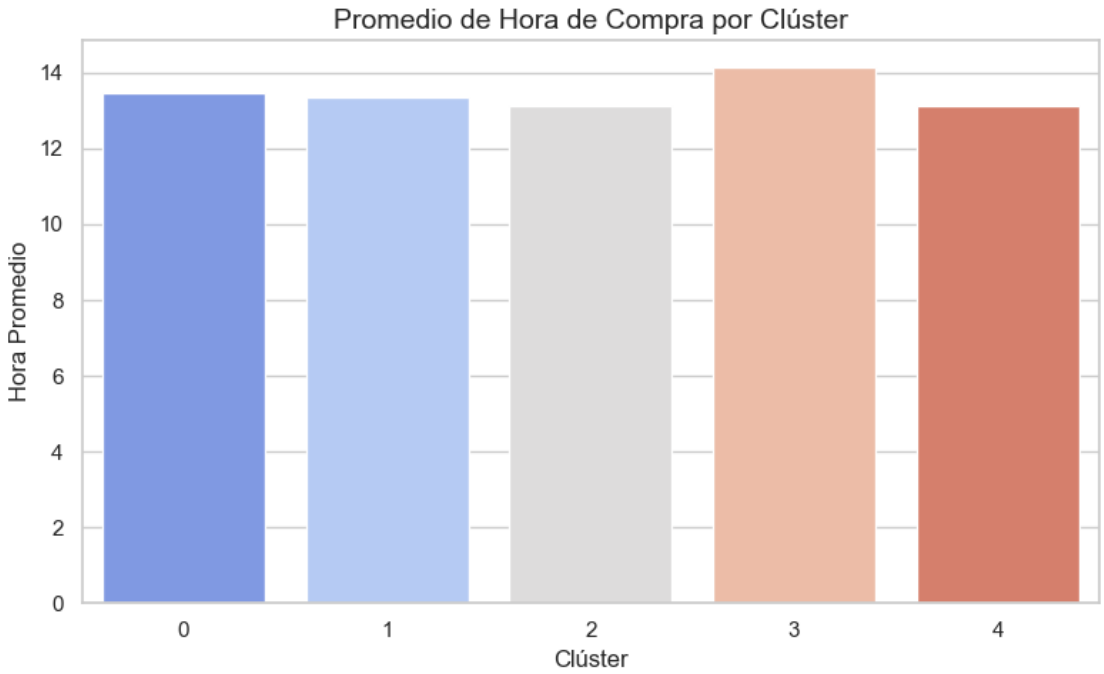

Figura 10. Promedio de hora de compra por clúster

Todos los grupos compran preferentemente entre las 13:00 y 14:00 hrs, lo cual podría coincidir con horas postalmuerzo o salidas laborales. Sin embargo, el clúster 3 destaca por tener el valor más alto (≈14.1 hrs).

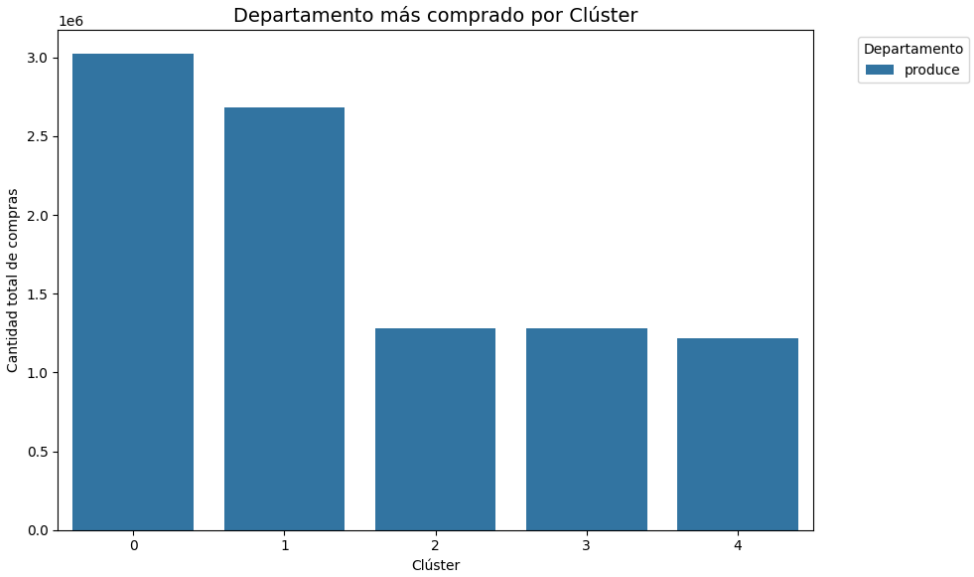

Figura 11. Departamento más comprado

El producto más comprado en todos los clústeres es *"produce"* (frutas y verduras), con gran diferencia. Esto sugiere que es un eje transversal de consumo saludable.

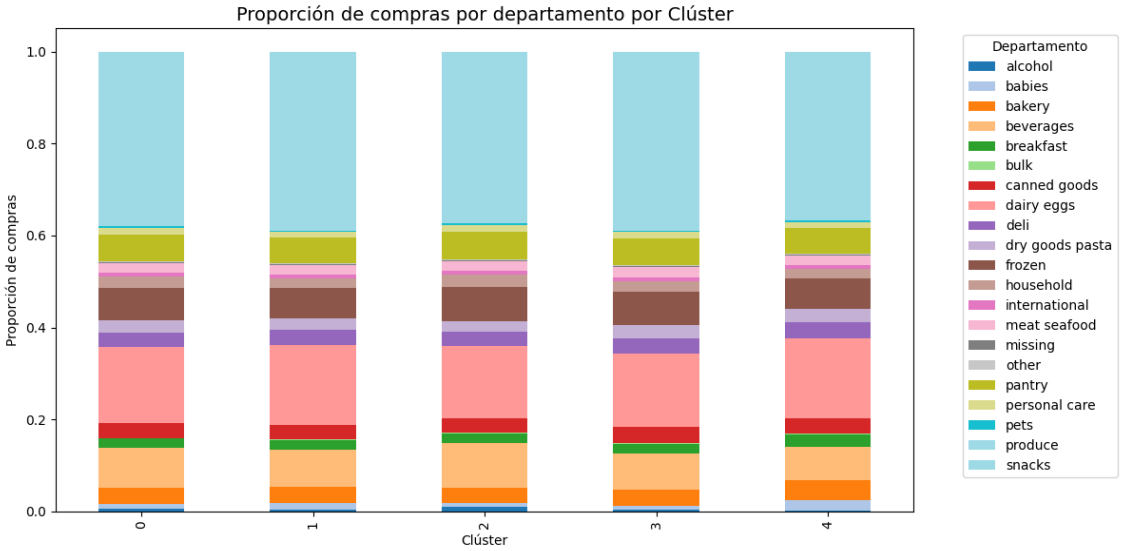

Figura 12. Proporción de compras por departamento

Si bien hay variación, la proporción de *“produce”*, *“dairy eggs”* y *“snacks”* es alta en todos los grupos. Algunos clústeres tienen compras más diversificadas, lo que puede guiar campañas segmentadas.

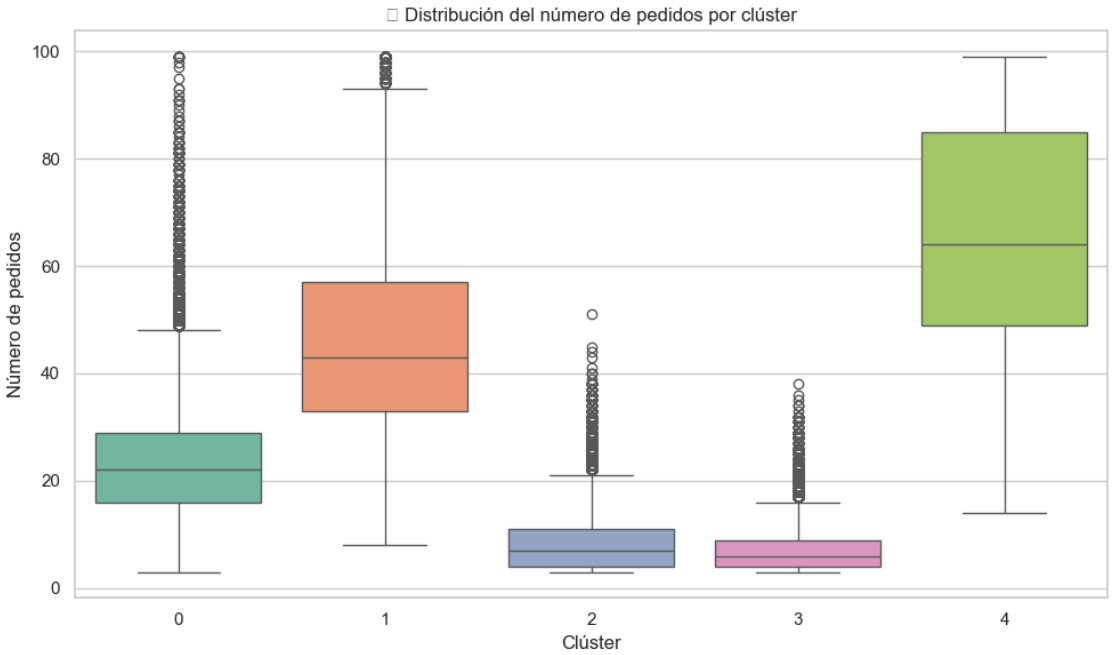

Figura 13. Distribución del número de pedidos por clúster

El clúster 4 sobresale con la mediana más alta de pedidos, mientras que los clústeres 2 y 3 tienen una frecuencia significativamente menor. Este dato puede orientar promociones de fidelización o beneficios por recurrencia.

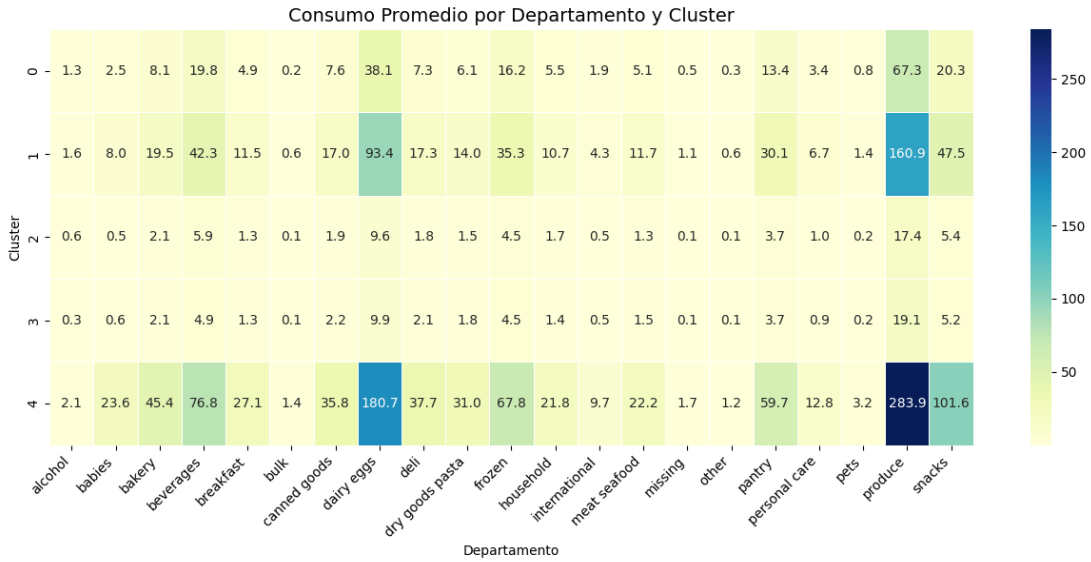

Figura 14. Consumo promedio por departamento y clúster

Se identificaron preferencias notables:

- Clúster 4: consumo excesivamente alto de *produce*, *dairy eggs*, *breakfast*, *snacks*.
- Clúster 1: consumo muy marcado en *dairy eggs* y *produce*.
- Clústeres 2 y 3: patrones más livianos, concentrados en pocos departamentos.

3- Hallazgos preliminares

El análisis exploratorio realizado mediante reglas de asociación y técnicas de clustering ha permitido identificar patrones significativos de comportamiento de los clientes en la base de datos de compras. A continuación, se resumen los hallazgos más relevantes:

Reglas de Asociación (A Priori)

1. **Bajo soporte, confianza y alto lift.** Las reglas extraídas muestran generalmente un bajo soporte y confianza, lo cual es esperable considerando la diversidad y cantidad de productos disponibles. Sin embargo, se han identificado reglas con altos valores de lift, lo que indica una fuerte dependencia entre ciertos productos, incluso si su frecuencia de aparición conjunta es baja. Esto es crucial para generar recomendaciones que sorprendan y generen valor.

2. **Agrupación en redes**. Al visualizar las reglas mediante grafos, se observa un núcleo de productos orgánicos que se relacionan estrechamente entre sí. Destacan productos como *Organic Strawberries*, *Organic Bananas* y *Organic Avocado*, los cuales actúan como nodos centrales. Esto sugiere que estos productos podrían servir como puntos de entrada para campañas de marketing cruzado.

Segmentación de Clientes (Clustering)

1. **Distribución heterogénea de clientes**. El modelo de K-Means con 5 clústeres reveló una distribución desequilibrada de clientes, donde los clústeres 2 y 3 concentran la mayoría de usuarios. Esto indica que hay segmentos predominantes de comportamiento de compra, lo que puede facilitar estrategias de segmentación.

2. **Diferencias en el número de pedidos.** El análisis boxplot muestra que ciertos clústeres, como el clúster 4, concentran a clientes muy activos, con una mayor cantidad de pedidos acumulados. En contraste, otros grupos

presentan una baja recurrencia de compra, lo cual podría estar asociado a compradores esporádicos o nuevos.

3. **Hora promedio de compra**. Los distintos grupos presentan variaciones horarias significativas en los hábitos de compra. Algunos clientes tienden a comprar por la mañana, mientras que otros lo hacen por la tarde. Esta información puede ser útil para programar notificaciones o promociones en horarios estratégicos.

4. **Preferencias por departamentos**. En todos los clústeres, el departamento "produce" (frutas y verduras) lidera ampliamente en volumen y proporción de compras, lo cual se refleja en múltiples visualizaciones. No obstante, se observaron diferencias notables en el consumo de categorías secundarias como *dairy eggs*, *snacks* o *bakery*, lo que permite definir perfiles de clientes con mayor granularidad.

5. **Heatmap de consumo por clúster**. El análisis detallado de consumo promedio por departamento refuerza que algunos segmentos tienen fijaciones claras por ciertos productos. Por ejemplo, el clúster 4 muestra una marcada tendencia a consumir productos del tipo *dairy eggs* y *produce*, lo cual puede ser clave para diseñar combos personalizados o promociones específicas.

4- Conclusión final del informe

Este estudio ha permitido evidenciar el poder de la analítica de datos aplicada al comportamiento de compra. A través de técnicas de minería de reglas de asociación y segmentación de clientes, fue posible descubrir patrones valiosos que de otro modo pasarían desapercibidos.

La utilización del algoritmo A Priori reveló relaciones significativas entre productos, incluso cuando estos no son comprados con alta frecuencia, lo cual representa una oportunidad directa para estrategias de recomendación y ventas cruzadas.

Por su parte, el análisis de clustering permitió detectar segmentos de usuarios con comportamientos diferenciados, facilitando una comprensión más fina del público objetivo. Este tipo de segmentación puede ser la base de campañas más efectivas, recomendaciones personalizadas, mejoras en la logística y optimización de la atención al cliente.

En definitiva, este trabajo demuestra cómo el uso integrado de herramientas como Python, Pandas, Scikit-learn y visualización puede transformar datos masivos en acciones concretas y estratégicas para las organizaciones. Los hallazgos presentados constituyen una base sólida para continuar con el desarrollo de dashboards interactivos, recomendaciones inteligentes y decisiones orientadas al cliente.

CAPÍTULO V: VISUALIZACIÓN DINÁMICA DE RESULTADOS MEDIANTE DASHBOARD INTERACTIVO

Como complemento visual e interactivo al análisis de datos realizado, se desarrolló un dashboard web utilizando la librería Dash de Plotly en Python. Este panel permite explorar dinámicamente los patrones de comportamiento de los clientes y las reglas de asociación descubiertas, facilitando así una comprensión más profunda para la toma de decisiones.

Elementos incluidos en el dashboard

1- Distribución de clientes por clúster. Permite observar el tamaño relativo de cada grupo detectado mediante K-means.

2- Hora promedio de compra por clúster. Muestra a qué hora suelen comprar los distintos segmentos de usuarios.

3- Consumo promedio por departamento y clúster. Un mapa de calor que evidencia qué departamentos son más o menos consumidos por cada segmento.

4- Reglas de asociación (soporte vs. confianza). Visualiza las reglas descubiertas por el algoritmo Apriori, resaltando aquellas con mayor lift.

5- Red interactiva de reglas de asociación. Un grafo que representa la relación entre productos comprados en conjunto, permitiendo detectar núcleos de afinidad.

Para abrir el dashboard localmente, se puede utilizar el archivo lanzar_dashboard.bat ubicado en la raíz del proyecto. Una vez ejecutado, el panel estará disponible en:

http://127.0.0.1:8050/

Figura 15. Dashboard interactivo 1

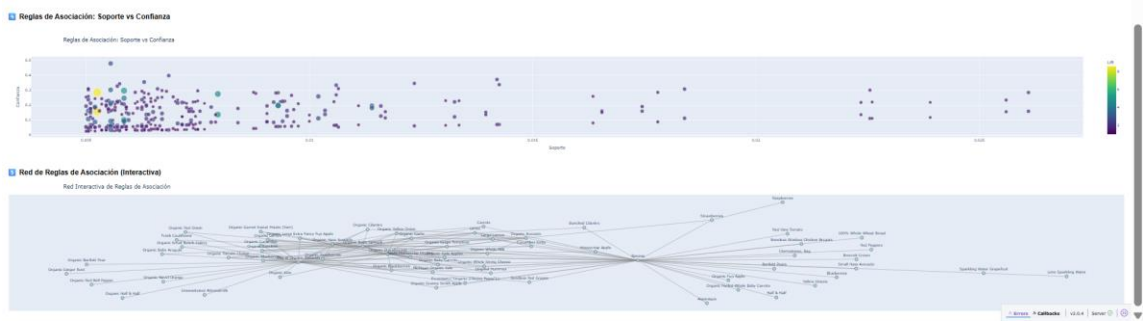

Figura 16. Dashboard interactivo 2

Se puede igualmente encontrar en el siguiente enlace de GitHub:

https://github.com/alfonso-abbott/portafolio_de_proyecto

REFERENCIAS

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207–216. https://doi.org/10.1145/170036.170072

Cámara de Comercio de Santiago. (2024). Informe de Comercio Electrónico en Chile 2023. Recuperado de https://www.ccs.cl

González, M., Pérez, A., & Vargas, L. (2023). Uso de técnicas de minería de datos para la optimización de ventas en supermercados online en Chile. Revista Latinoamericana de Comercio Electrónico, 18(2), 45–62.

Grover, P., Kar, A. K., & Ilavarasan, P. V. (2020). Impact of big data analytics on e-commerce and retail firms: A systematic review. Information Systems Frontiers, 22(5), 1279–1300. https://doi.org/10.1007/s10796-019-09926-2

Ramesh, V., & Baskar, D. (2021). Comparative analysis of clustering algorithms for customer segmentation in retail industry. International Journal of Engineering Research and Technology, 10(3), 456–463. https://doi.org/10.17577/IJERTV10IS030257

Song, H., & Kim, S. (2022). Customer purchase behavior analysis using data mining techniques in online shopping malls. Journal of Retailing and Consumer Services, 64, 102759. https://doi.org/10.1016/j.jretconser.2021.102759