FACULTAD DE INGENIERÍA
INGENIERÍA EN COMPUTACIÓN E INFORMÁTICA

ANÁLISIS DE PATRONES DE COMPRA Y SEGMENTACIÓN DE CLIENTES, CASO LIDER.CL MEDIANTE MINERÍA DE DATOS

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 noviembre de 2025, en la ciudad de Santiago.

Facultad de Ingeniería

Escuela Computación e Informática

Título del trabajo Análisis de patrones de compra y Segmentación de clientes, caso lider.cl mediante minería de datos

Nombre y firma del autor

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN EJECUTIVO

Este proyecto desarrolla una solución analítica orientada a identificar patrones de compra y segmentar clientes en el contexto de un supermercado online. Se aplican técnicas descriptivas de minería de datos, particularmente reglas de asociación (Apriori) y clustering (K-means), para analizar datos transaccionales simulados y generar perfiles de comportamiento.

El enfoque metodológico se basa en CRISP-DM, complementado con una gestión híbrida entre Scrum y PMBOK. Los resultados se presentan mediante visualizaciones interactivas en un dashboard desarrollado con Dash, incluyendo además un cruce entre las reglas y los clústeres obtenidos.

El desarrollo se realizó en Python (VS Code), con librerías especializadas para procesamiento, modelado y visualización. Se entrega una solución replicable, documentada y aplicable a contextos reales de análisis comercial.

PALABRAS CLAVE

ciencia de datos, reglas de asociación, K-means, A priori, segmentación de clientes, minería de datos, dashboard interactivo, patrones de compra, Python, retail online, CRISP-DM

ABSTRACT

This project develops an analytical solution aimed at identifying purchasing patterns and segmenting customers in the context of an online supermarket. Descriptive data mining techniques—specifically association rules (Apriori) and clustering (K-means)—are applied to simulated transactional data to generate behavioral profiles.

The methodology follows the CRISP-DM model, complemented by a hybrid management approach combining Scrum and PMBOK. Results are presented through interactive visualizations in a Dash-based dashboard, including a cross-analysis between rules and clusters.

The entire development was carried out in Python (VS Code), using specialized libraries for processing, modeling, and visualization. The final product is a documented, replicable solution, applicable to real-world commercial analysis.

KEY WORDS

data science, association rules, K-means, A priori, customer segmentation, data mining, interactive dashboard, purchase patterns, Python, online retail, CRISP-DM

CAPÍTULO I: INTRODUCCIÓN

1- Introducción

En la actualidad, los supermercados online generan grandes volúmenes de datos a partir de transacciones, hábitos de compra y preferencias de sus clientes. Sin embargo, en muchos casos estos datos permanecen subutilizados, limitando su potencial para optimizar estrategias comerciales, personalizar recomendaciones o anticipar demandas del mercado.

El presente proyecto surge como respuesta a esta oportunidad, desarrollando una solución basada en ciencia de datos que permite descubrir patrones ocultos en los datos transaccionales y segmentar a los clientes según su comportamiento de compra. La propuesta se enmarca dentro del dominio de la minería de datos descriptiva, utilizando técnicas como las reglas de asociación (Apriori) y el análisis de clústeres (K-means), con el objetivo de generar insights que puedan traducirse en decisiones informadas y acciones comerciales estratégicas.

La metodología adoptada para el desarrollo es CRISP-DM, ampliamente validada en el campo del análisis de datos, y se complementa con una gestión de proyecto híbrida que combina elementos de Scrum y el marco PMBOK. Esta estructura metodológica permite un avance iterativo, flexible y riguroso, adecuado tanto para el contexto académico como para escenarios reales de aplicación.

El desarrollo técnico se realiza en Python, utilizando entornos como Visual Studio Code y librerías especializadas en procesamiento de datos, modelado y visualización. Los resultados del análisis son presentados mediante visualizaciones interactivas integradas en un dashboard funcional, lo que permite comunicar de

forma clara los hallazgos y facilitar su interpretación por parte de usuarios no técnicos.

Finalmente, el proyecto incluye un cruce exploratorio entre los resultados de segmentación y las reglas de asociación obtenidas, con el fin de enriquecer la comprensión de los distintos perfiles de clientes. Este enfoque integrado representa un aporte metodológico y práctico para el análisis de comportamiento en entornos de retail online.

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, donde el conocimiento profundo del comportamiento de sus clientes se vuelve un factor clave para sostener ventajas estratégicas. Supermercados como Lider.cl, que manejan un amplio catálogo de productos y un volumen masivo de transacciones, requieren enfoques analíticos avanzados para diferenciarse, fidelizar a sus clientes y maximizar sus ingresos. Sin embargo, muchas decisiones comerciales aún se toman sin considerar el potencial de la información contenida en los datos transaccionales.

La capacidad de detectar patrones en el comportamiento de compra permite responder preguntas como: ¿qué productos suelen adquirirse conjuntamente?, ¿existen segmentos de clientes con preferencias similares?, ¿en qué horarios o días se concentran determinadas categorías de productos? Abordar estas interrogantes mediante herramientas de minería de datos, enmarcadas dentro de un enfoque más amplio de ciencia de datos aplicada, tiene un alto valor estratégico. Permite mejorar la experiencia del cliente mediante recomendaciones más precisas, promociones

mejor focalizadas y una organización de productos más coherente con los hábitos reales de consumo.

No contar con estos mecanismos analíticos genera desventajas competitivas importantes. Las empresas que no logran identificar los patrones de comportamiento de sus clientes tienden a diseñar campañas genéricas, desperdiciar recursos publicitarios y desaprovechar oportunidades de ventas cruzadas. A su vez, el desconocimiento de la segmentación real de la clientela limita la posibilidad de establecer vínculos personalizados y sostenibles en el tiempo.
En el caso de supermercados online como Lider.cl, donde la oferta de productos es extensa y la frecuencia de compra elevada, la ausencia de análisis automatizado puede derivar en ineficiencias logísticas, sobrestock de productos poco demandados y disminución en la fidelización de segmentos clave. Implementar mecanismos de análisis que permitan explotar los datos disponibles dejó de ser una ventaja competitiva: se ha transformado en una necesidad estratégica.

Este proyecto aborda directamente esta problemática, desarrollando una solución basada en técnicas de minería de datos para descubrir asociaciones relevantes entre productos frecuentemente adquiridos juntos, segmentar a los clientes en grupos homogéneos y, en una etapa posterior, cruzar ambas dimensiones para generar conocimiento más granular sobre los perfiles de consumo y sus preferencias. Esta estrategia entrega información valiosa para apoyar decisiones como la ubicación de productos en la plataforma, la generación de recomendaciones automáticas y la planificación de campañas promocionales personalizadas. El análisis desarrollado tiene un impacto directo sobre indicadores clave de rendimiento comercial, tales como la tasa de conversión, el valor promedio del carrito y la frecuencia de recompra.

3- Breve discusión bibliográfica

La minería de datos ha demostrado ser una herramienta clave para extraer valor estratégico desde grandes volúmenes de información transaccional, especialmente en entornos de e-commerce. Las reglas de asociación, como las generadas mediante el algoritmo Apriori, permiten identificar relaciones significativas entre productos comprados en conjunto, habilitando estrategias de ventas cruzadas, optimización del ticket promedio y mejoras en la disposición de productos tanto físicos como digitales.

Diversos estudios respaldan el valor de estas técnicas. Por ejemplo, Grover, Kar e Ilavarasan (2020) destacan cómo el descubrimiento de patrones de consumo permite personalizar campañas, incrementar la fidelización y potenciar la eficiencia comercial en el retail digital. Paralelamente, el análisis del comportamiento de compra por segmentos ha sido ampliamente abordado mediante técnicas de clustering, siendo K-means una de las más utilizadas por su simplicidad, capacidad de escalabilidad y claridad interpretativa.

Un caso particularmente relevante es el presentado por Song y Kim (2022), quienes aplicaron reglas de asociación y clustering de clientes en una plataforma de e-commerce asiática. La combinación de ambas técnicas logró mejorar en un 25% la tasa de conversión, validando empíricamente el valor de aplicar un enfoque híbrido, como el propuesto en este proyecto.

En esta misma línea, Ramesh y Baskar (2021) realizaron una comparación entre métodos de clustering aplicados al retail, concluyendo que K-means permite agrupar de forma efectiva a los clientes en segmentos homogéneos, facilitando decisiones de marketing personalizadas y optimización del inventario.

Finalmente, desde una mirada regional, González et al. (2023) analizaron el estado del análisis de datos en supermercados online chilenos, evidenciando que, a pesar del crecimiento del e-commerce, muchas empresas aún no aprovechan técnicas avanzadas como Apriori o K-means. Esta brecha representa no solo una debilidad competitiva, sino también una oportunidad estratégica para incorporar procesos analíticos en la transformación digital del sector.

En conjunto, estos estudios sustentan tanto la validez técnica como la pertinencia práctica del enfoque utilizado en este proyecto: no solo se aplican dos técnicas consolidadas en el análisis del comportamiento de compra, sino que se propone una articulación metodológica novedosa —el cruce entre Apriori y K-means— que permite generar insights más precisos, aplicables directamente a la personalización en plataformas como Lider.cl.

4- Contribución del trabajo

La contribución principal de este proyecto se enmarca en el ámbito de la ciencia de datos, con un foco específico en la minería de datos aplicada al comercio electrónico. Se propone un análisis de patrones de compra y segmentación de clientes orientado a supermercados online, utilizando como caso de referencia el contexto operativo de Lider.cl. Esta contribución se articula en distintos niveles:

En primer lugar, se genera conocimiento estratégico a partir de datos transaccionales reales. A través de reglas de asociación, se identifican combinaciones frecuentes de productos comprados conjuntamente, lo que permite diseñar estrategias de ventas cruzadas, paquetes promocionales y campañas publicitarias más alineadas con el comportamiento del cliente.

En segundo lugar, mediante la aplicación del algoritmo K-means, se realiza una segmentación de clientes que agrupa consumidores con comportamientos de compra similares. Esta segmentación facilita la personalización de acciones de marketing, la optimización del inventario y una mayor fidelización a través de propuestas adaptadas a cada perfil.

Como tercer aporte, se incorpora un enfoque de análisis combinado: el cruce de reglas de asociación y clustering. Esta estrategia permite identificar reglas diferenciadas por tipo de cliente, mejorando la precisión de las recomendaciones y revelando asociaciones significativas dentro de cada grupo de consumo. Este enfoque aporta valor agregado al análisis tradicional, habilitando decisiones más informadas y específicas.

Desde una perspectiva tecnológica, el proyecto implementa sus análisis mediante herramientas de código abierto como Python, utilizando librerías especializadas (mlxtend, pandas, scikit-learn, entre otras). Este enfoque garantiza reproducibilidad, escalabilidad y adaptabilidad del proceso a otras plataformas de retail digital.

Además, se fortalece la cultura organizacional basada en la toma de decisiones informadas por datos, al demostrar cómo el uso de técnicas avanzadas de minería de datos puede integrarse a los flujos operacionales y estratégicos de empresas del sector.

Finalmente, como resultado tangible, se ha construido un dashboard interactivo con visualizaciones dinámicas en Plotly y Dash, lo cual amplía el alcance práctico del análisis al ofrecer una herramienta accesible, exploratoria y potencialmente integrable en entornos reales de negocio.

5- Trabajo realizado en el proyecto

El presente proyecto desarrolla un análisis de datos transaccionales basado en un conjunto de datos públicos representativos del comercio electrónico de supermercado, simulando su aplicación en el contexto chileno de Lider.cl. El propósito fue identificar patrones de comportamiento de compra y segmentar clientes con el fin de apoyar decisiones estratégicas en ventas, marketing y fidelización.

El trabajo realizado abarcó las siguientes etapas:

- Recopilación y preprocesamiento de datos: Se seleccionó el dataset *Instacart Online Grocery Shopping 2017*, el cual fue transformado para representar el entorno operativo de un supermercado online. Se integraron archivos de productos, pedidos, departamentos y categorías, y se construyeron tablas de transacciones y perfiles de clientes mediante scripts en Python.

- Análisis de reglas de asociación: Se aplicó el algoritmo Apriori para descubrir combinaciones frecuentes de productos adquiridos conjuntamente. Se generaron métricas relevantes como soporte, confianza y lift, y se visualizó el conjunto de reglas tanto en gráficos de dispersión como en redes interactivas.

- Segmentación de clientes por clustering: Se implementó el algoritmo K-means, utilizando validaciones mediante el método del codo y el coeficiente de silueta. Se construyeron perfiles de clientes según variables como número de pedidos, horario promedio de compra y consumo por departamento.

- Cruce entre reglas de asociación y clusters: Se diseñó una metodología para combinar los resultados de Apriori y K-means, obteniendo reglas de asociación diferenciadas por clúster. Esto permitió identificar asociaciones relevantes según los perfiles detectados, agregando un nivel de personalización al análisis.

- Desarrollo de dashboard interactivo: Se construyó una interfaz web utilizando Dash y Plotly, en la cual se integran las visualizaciones generadas en cada etapa. Esta herramienta permite la exploración dinámica de los clústeres, reglas de asociación y combinaciones personalizadas por segmento de cliente.

- Documentación y análisis de hallazgos: Todos los resultados fueron interpretados y presentados mediante gráficos, tablas y explicaciones técnicas, estructurados en un informe reproducible y comprensible para públicos no especializados.

Las herramientas principales utilizadas fueron Python (con librerías como pandas, mlxtend, scikit-learn, plotly, dash, networkx) y R (para exploración alternativa). Todo el trabajo se estructuró en scripts modulares, almacenados y versionados según buenas prácticas de desarrollo.

CAPÍTULO II: IDENTIFICACIÓN DEL PROBLEMA / OPORTUNIDAD

1- Presentación y fundamentación del problema

El comercio electrónico ha redefinido la relación entre consumidores y empresas, estableciendo nuevos estándares de conveniencia y personalización. En este contexto, el sector supermercadista ha acelerado su digitalización. Plataformas como Lider.cl, de Walmart Chile, han consolidado su presencia online, pero enfrentan desafíos para mantener la competitividad, especialmente en lo que respecta a la comprensión del comportamiento de sus clientes.

Pese a contar con vastas bases de datos —transacciones, navegación, horarios y productos—, muchas decisiones comerciales siguen basándose en intuiciones o segmentaciones generales. Esto se traduce en campañas poco relevantes, baja conversión y una experiencia de compra genérica, desaprovechando el potencial de los datos disponibles.

La ausencia de personalización efectiva afecta áreas clave: promociones, logística, fidelización y planificación comercial. Por ejemplo, sin conocer los productos que suelen adquirirse juntos, se pierden oportunidades para aplicar ventas cruzadas inteligentes. Sin segmentación basada en hábitos reales, se dificulta la asignación óptima de recursos.

Desde una mirada estratégica, esta brecha entre disponibilidad de datos y su aprovechamiento impide avanzar hacia una cultura de decisiones basada en evidencia. En un entorno digital competitivo, integrar analítica avanzada es esencial para la sostenibilidad.

Este proyecto plantea una oportunidad concreta: desarrollar un sistema analítico que permita explorar datos transaccionales mediante minería de datos. Se simula su aplicación en Lider.cl utilizando un dataset público del retail norteamericano, con tres líneas de análisis:

- Apriori: para descubrir productos comprados juntos y optimizar recomendaciones y promociones.
- K-means: para segmentar clientes según su comportamiento de compra y permitir campañas más efectivas.
- Cruce entre ambos: para identificar combinaciones de productos específicas por segmento, habilitando estrategias aún más personalizadas.

En síntesis, el problema reside en el bajo aprovechamiento del potencial analítico de los datos disponibles. Este proyecto propone una solución integrada que transforma datos en decisiones estratégicas, sentando las bases para futuras aplicaciones como motores de recomendación, predicción de demanda o personalización dinámica.

2- Descripción de problemas / oportunidades de mejora

A partir del diagnóstico del entorno digital del retail, se identifican diversos problemas que, correctamente abordados mediante técnicas analíticas, pueden transformarse en oportunidades de mejora.

Problemas identificados

1. Falta de personalización: La navegación y recomendaciones no se ajustan al historial ni patrones reales del cliente, generando una experiencia genérica y baja conversión.

2. Marketing no segmentado: Las campañas se envían sin distinción entre tipos de consumidores, reduciendo su efectividad y aumentando el costo por adquisición.
3. Ausencia de ventas cruzadas inteligentes: Al no conocer productos comprados en conjunto, se desaprovechan oportunidades de maximizar el ticket promedio.
4. Gestión logística reactiva: La falta de análisis asociativo limita la planificación de demanda y genera sobrestock o quiebres.
5. Baja fidelización: La relación con el cliente no se basa en datos, dificultando la retención frente a competidores más personalizados.

Oportunidades de mejora

1. Minería de asociaciones (Apriori): Permite identificar productos complementarios, optimizar promociones y mejorar la conversión.
2. Segmentación por comportamiento (K-means): Agrupa clientes según hábitos reales, habilitando campañas y experiencias diferenciadas.
3. Combos personalizados por clúster: Combina reglas y segmentos para diseñar ofertas basadas en la lógica real de consumo.
4. Optimización de stock: Anticipa patrones de compra conjunta, mejorando la planificación logística y reduciendo pérdidas.
5. Impulso a la analítica organizacional: Integra procesos avanzados que fortalecen la madurez digital y la toma de decisiones basada en evidencia.

3- Identificación cuantitativa (ishikawa / árbol de oportunidades)

Introducción y justificación del uso de herramientas

Para analizar de forma estructurada la falta de análisis automatizado de patrones de compra y segmentación de clientes en Lider.cl, se utilizaron dos herramientas complementarias: el Diagrama de Ishikawa y el Árbol de Oportunidades. El primero permite descomponer el problema en causas organizadas por áreas clave (personas, métodos, tecnología, información, gestión y medición), mientras que el segundo proyecta los efectos positivos esperables al intervenir dichas causas mediante técnicas de minería de datos.

Ambas herramientas son pertinentes para un entorno e-commerce, ya que facilitan la visualización sistémica del problema y orientan la toma de decisiones estratégicas basadas en datos.

Descripción de causas según Diagrama de Ishikawa

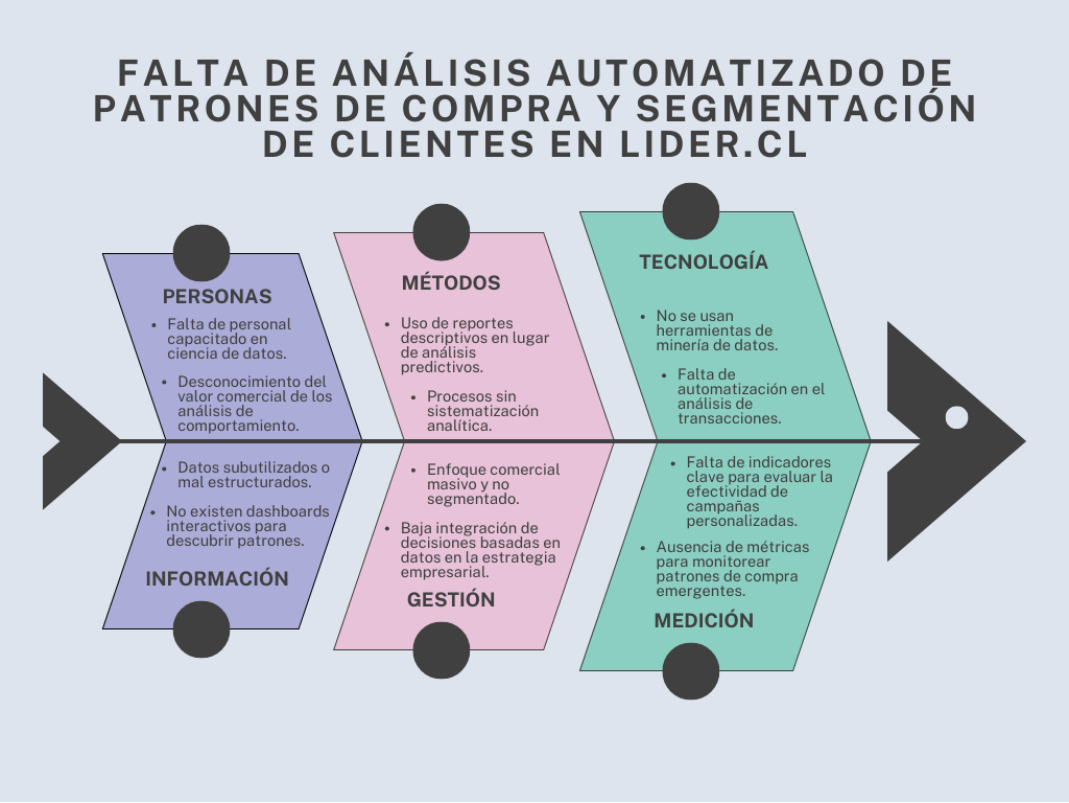

Figura 1: Diagrama de Ishikawa.

Las causas identificadas se agrupan en seis dimensiones críticas:

- Personas: ausencia de especialistas en ciencia de datos y desconocimiento del potencial analítico en áreas comerciales.
- Métodos: predominio de reportes descriptivos y falta de metodologías analíticas formalizadas.
- Tecnología: inexistencia de herramientas de minería de datos y procesos manuales que limitan la automatización.
- Información: datos subutilizados y no estructurados para análisis avanzado; historial transaccional desaprovechado.
- Gestión: decisiones basadas en intuición y enfoque operativo que desplaza iniciativas estratégicas.
- Medición: ausencia de métricas centradas en comportamiento de compra y falta de retroalimentación analítica.

Análisis del Árbol de Oportunidades

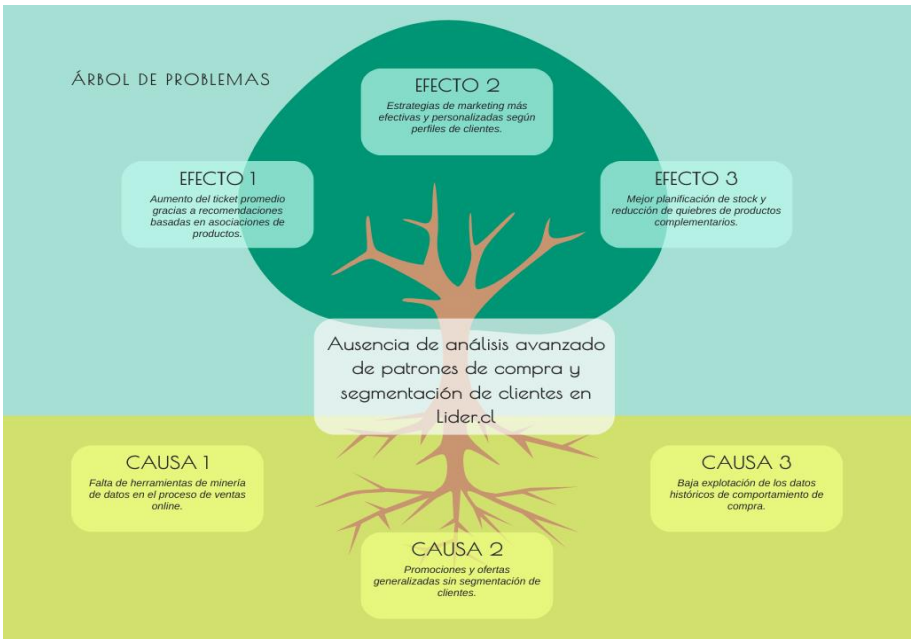

Figura 2: Árbol de Oportunidades.

El árbol permite visualizar la relación entre causas, problema central y beneficios.

- Raíces (causas estructurales): falta de herramientas analíticas, bajo uso del historial de compras, estrategias no personalizadas y desconocimiento organizacional sobre minería de datos.
- Tronco (problema central): ausencia de análisis automatizado de patrones de compra y segmentación.
- Copa (efectos esperados): aumento del ticket promedio, mejor segmentación y fidelización, personalización de campañas, optimización de inventario y fortalecimiento de la cultura de decisiones basadas en datos.

4- Objetivo general

Desarrollar un sistema de análisis basado en ciencia de datos y técnicas de minería de datos que permita identificar patrones de compra y segmentar clientes en un entorno simulado representativo del supermercado online Lider.cl, integrando el cruce de reglas de asociación (Apriori) con agrupamiento de clientes (K-means), 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

A partir del objetivo general, se definen los siguientes objetivos específicos, cada uno acompañado de su respectiva métrica de evaluación y un criterio de éxito que permite validar su cumplimiento:

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%.

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.

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.

4- Cruzar los resultados de segmentación y asociación para generar recomendaciones personalizadas por perfil de cliente.

- Métrica: Número de combinaciones relevantes entre clústeres y reglas de compra.
- Criterio de éxito: Al menos 3 perfiles de clientes con recomendaciones derivadas del cruce Apriori–K-means.

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.
Cruzar los resultados de segmentación y asociación para generar recomendaciones por perfil	Número de combinaciones entre clúster y reglas.	≥ 3 perfiles con recomendaciones específicas.

Tabla 1: Métricas de los objetivos específicos.

7- Limitaciones y alcances del proyecto

Alcances

El proyecto desarrolla una solución analítica que simula la aplicación de técnicas de minería de datos en un contexto similar al de Lider.cl, utilizando un dataset transaccional público. El enfoque integra reglas de asociación (Apriori), segmentación de clientes mediante K-means, un cruce analítico entre ambos resultados y la construcción de un dashboard interactivo orientado a la toma de decisiones.

Las actividades consideradas incluyen:

- Preparación del dataset: limpieza, transformación, codificación y estandarización.
- Extracción de reglas de asociación para identificar productos comprados conjuntamente.

- Segmentación de clientes con K-means, basándose en patrones de comportamiento de compra.
- Cruce Apriori–K-means para generar insights accionables diferenciados por clúster.
- Diseño de un dashboard con Dash + Plotly, con visualizaciones dinámicas, filtros y componentes interactivos.
- Elaboración de un informe técnico, que documenta la metodología, los hallazgos y las recomendaciones comerciales.

Limitaciones

- Uso de datos no reales: el análisis se realiza con datasets públicos debido a restricciones de acceso. Esto limita la personalización de los resultados, aunque no su validez metodológica.
- Enfoque académico–exploratorio: no contempla integración con sistemas corporativos, motores de recomendación operativos ni despliegue en entornos productivos.
- Restricciones técnicas y temporales: el trabajo se ejecuta en un entorno local y en un tiempo acotado; no se emplean servicios cloud, bases de datos online ni herramientas propietarias.
- Alcance de la generalización: las conclusiones dependen del dataset analizado. Su aplicación en entornos reales requiere validación adicional y ajustes según la industria o contexto comercial.

8- Normativa y leyes asociadas al proyecto

Este proyecto se desarrolla en un contexto académico utilizando datos públicos, anónimos y no vinculados a personas reales. Por lo tanto, no implica tratamiento de datos personales y no se ve afectado por la normativa vigente en materia de privacidad y protección de datos.

Sin embargo, una implementación real del sistema en una plataforma como Lider.cl sí involucraría datos personales y, por tanto, estaría sujeta al marco regulatorio correspondiente. En ese escenario, sería obligatorio cumplir con:

- Ley N.° 19.628 sobre Protección de la Vida Privada (Chile): regula el tratamiento de datos personales, exige consentimiento informado, establece principios de finalidad, proporcionalidad y seguridad, y obliga al responsable del tratamiento a resguardar la confidencialidad y correcta utilización de la información.

- Políticas internas de privacidad y términos de uso de Lider.cl: cualquier análisis de comportamiento debe alinearse con los consentimientos otorgados por los usuarios y con las restricciones sobre uso de datos para fines analíticos, comerciales o automatizados.

Dado que este proyecto utiliza datos simulados y no identifica a individuos, no incurre en riesgos legales. No obstante, una futura adopción en un entorno corporativo requeriría un cumplimiento estricto de la ley, de las políticas internas y de estándares éticos en el manejo de información sensible.

CAPÍTULO III: MARCO METODOLÓGICO

1- Metodología de desarrollo

Este proyecto adopta una estrategia metodológica dual que articula dos enfoques complementarios:

- Una metodología de desarrollo analítico centrada en el uso de minería de datos.
- Un modelo de gestión híbrido, que combina agilidad y control estructurado para organizar y monitorear el proyecto.

1.1. Metodología de desarrollo: CRISP-DM

El análisis de datos se desarrolla bajo el enfoque CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente utilizado en ciencia de datos por su flexibilidad y estructura iterativa. Sus seis fases guían el proceso desde el planteamiento del problema hasta la implementación de resultados:

1. Comprensión del negocio: Se identifican los desafíos del entorno digital de supermercados como Lider.cl y se traducen en objetivos analíticos concretos.
2. Comprensión de los datos: Se selecciona un dataset público representativo del comercio electrónico y se analiza su estructura y calidad.
3. Preparación de los datos: Incluye limpieza, transformación y codificación de variables para asegurar su usabilidad.
4. Modelado: Se aplican reglas de asociación (Apriori) y segmentación (K-means), con un cruce final que relaciona productos y perfiles de cliente.
5. Evaluación: Se analizan métricas clave como soporte, confianza y Silhouette Score, junto con validaciones visuales y estratégicas.

6. Despliegue: Se entregan los hallazgos en un informe técnico y se implementa un dashboard interactivo con Plotly y Dash que simula su uso en un sistema real.

Este enfoque permite un desarrollo sistemático, con foco en la calidad de los datos y la utilidad estratégica de los resultados.

1.2. Metodología de gestión: híbrida (Scrum + PMBOK)

Para organizar el trabajo, se adopta una metodología híbrida que combina:

- Scrum: para dividir el desarrollo en iteraciones semanales (sprints), con tareas priorizadas como análisis, modelado y visualización.
- PMBOK: para estructurar elementos clave del proyecto como el alcance, los riesgos, la calidad y la documentación formal.

Esta combinación proporciona flexibilidad en la ejecución, sin perder trazabilidad ni control del progreso, lo cual es especialmente útil en proyectos de exploración analítica con plazos definidos.

1.3. Validación metodológica previa

El uso de CRISP-DM y del enfoque híbrido se sustenta en la experiencia adquirida en tres cursos formativos aplicados:

- IBM Ciencia de Datos
- IBM AI Developer
- IBM Minería de Datos (Skills Network)

En estos programas se abordaron casos reales con Python, se aplicaron modelos como Apriori y K-means, y se trabajó con dashboards interactivos, metodologías iterativas y documentación estructurada. Esta base permite aplicar las metodologías en este proyecto de forma profesional, informada y alineada con estándares de la industria.

2- Herramientas y ambiente de desarrollo

El proyecto se desarrolló utilizando herramientas propias de entornos profesionales de análisis de datos, seleccionadas por su eficiencia, flexibilidad y capacidad para integrar minería de datos y visualización interactiva. El entorno principal fue Visual Studio Code, que permitió organizar los scripts en carpetas modulares, gestionar versiones con Git y trabajar con un entorno virtual de Python. Aunque se realizaron pruebas iniciales en Google Colab, todo el desarrollo final —incluida la documentación y las visualizaciones— se ejecutó completamente en VS Code para asegurar control total del flujo de archivos.

Python fue el lenguaje central, apoyado en un archivo *requirements.txt* generado desde el entorno virtual, lo que garantiza la reproducción del proyecto mediante pip install -r requirements.txt. La organización de carpetas siguió las etapas del pipeline (preprocesamiento, minería, visualización e integración), mientras que gráficos y archivos CSV se almacenaron en directorios separados para facilitar la trazabilidad. Entre las librerías destacadas utilizadas se encuentran:

- pandas y numpy para manipulación de datos y estructuras numéricas.
- mlxtend y scikit-learn para la implementación de algoritmos de minería de datos como Apriori y K-means, respectivamente.
- matplotlib, seaborn y plotly para generación de gráficos estáticos e interactivos.

- dash y Flask para el desarrollo de un dashboard final interactivo como prototipo de solución web visual.
- deep-translator y beautifulsoup4 en funciones auxiliares, como procesamiento textual y scraping, aunque su uso fue marginal respecto al núcleo del proyecto.

En conjunto, estas herramientas hicieron posible construir un pipeline completo - desde la preparación de datos hasta una interfaz analítica funcional- sin depender de infraestructura externa, asegurando portabilidad y reproducibilidad en un contexto académico.

3- Descripción general de la propuesta de solución

La solución propuesta en este proyecto se estructura como un sistema de análisis de datos transaccionales, diseñado para identificar patrones de comportamiento de compra y segmentar a los clientes de un supermercado online mediante técnicas de minería de datos. El objetivo es simular cómo este tipo de análisis puede integrarse en una plataforma como Lider.cl para optimizar su estrategia comercial.

Este sistema se desarrolla de forma modular y su implementación se basa en el procesamiento, análisis y visualización de datos públicos que emulan el comportamiento real de compra en entornos e-commerce.

3.1 Enfoque general de la solución

El proyecto se divide en tres grandes etapas analíticas:

1. **Análisis de reglas de asociación (Apriori):** Se utiliza el algoritmo Apriori para detectar combinaciones de productos que suelen adquirirse juntos por

los clientes. Las reglas generadas se evalúan con métricas como soporte, confianza y lift, y se visualizan mediante gráficos de red, barras y mapas de calor. Esta etapa permite identificar oportunidades de venta cruzada, creación de combos o recomendaciones automáticas.

2. **Segmentación de clientes (Clustering con K-means):** A partir de atributos derivados del historial de compras (como frecuencia de pedidos, total gastado, variedad de departamentos comprados y hora promedio), se agrupa a los clientes en clústeres homogéneos utilizando el algoritmo K-means. Esta etapa incluye una validación rigurosa del modelo mediante análisis de correlación, método del codo e índice de silueta. La segmentación permite definir perfiles de cliente y aplicar estrategias diferenciadas por grupo.

3. **Cruce entre reglas de asociación y clústeres (Análisis cruzado Apriori-K-means):** Se realiza un cruce entre los resultados de ambas técnicas, permitiendo filtrar las reglas de asociación por clúster. Esta combinación mejora la personalización de las recomendaciones, ya que permite identificar productos frecuentemente comprados dentro de cada grupo específico de clientes.

3.2 Resultados intermedios como entregables

Durante el desarrollo del sistema se generaron múltiples productos analíticos, organizados en dos tipos de entregables:

- Archivos estructurados (.csv): que contienen salidas del modelo, como reglas filtradas, perfiles agregados por clúster, métricas de validación o resúmenes de comportamiento.

- Visualizaciones gráficas (.png): que permiten interpretar de forma clara los patrones descubiertos y validar el comportamiento de los algoritmos aplicados.

Algunos de los productos más relevantes incluyen:

- Gráficos de reglas de asociación (red amplia, red reducida, barras, mapa de calor).
- Visualizaciones del clustering (distribución por clúster, PCA, promedio de hora, número de pedidos).
- Análisis de preferencias por departamento y perfiles agregados por radar chart.
- Validación del modelo de K-means (heatmaps de correlación, método del codo, silueta por K).

Estos productos no solo constituyen evidencia del proceso analítico, sino que también son fácilmente integrables en reportes técnicos, presentaciones ejecutivas o sistemas de apoyo a la toma de decisiones.

3.3 Prototipo funcional: dashboard interactivo

Como exploración adicional, se construyó un dashboard interactivo utilizando la biblioteca Dash de Python. Este prototipo permite visualizar de forma dinámica los principales resultados del análisis, incluyendo las reglas descubiertas, la distribución de clústeres, los comportamientos agregados por segmento y la generación de recomendaciones personalizadas. Si bien no constituye un producto final listo para producción, cumple un rol ilustrativo sobre cómo estos análisis podrían ser desplegados en una plataforma comercial real.

3.4 Justificación de la solución

Este enfoque modular basado en Python, archivos .csv e imágenes de salida, prioriza la reproducibilidad, la trazabilidad del análisis y la claridad en la presentación de resultados. La solución no requiere infraestructura avanzada ni sistemas externos, lo cual la hace viable para entornos académicos o equipos de ciencia de datos en etapas exploratorias. Además, su estructura permite ser extendida o automatizada en futuras iteraciones, integrando motores de recomendación, dashboards en producción o conexión con bases de datos reales.

4- Propuesta de controles y evidencia

Durante el desarrollo del proyecto se aplicaron mecanismos para asegurar control, trazabilidad y coherencia en todo el flujo analítico, tanto en el procesamiento técnico de los datos como en la organización y respaldo del trabajo. Toda la evidencia generada fue almacenada y versionada de forma ordenada para permitir su verificación y reproducción futura.

Se definió una estructura de carpetas que separa scripts, datos procesados y visualizaciones, de modo que cada salida generada mantuviera correspondencia directa con el código fuente. En cada etapa se registraron las transformaciones aplicadas y se generaron archivos intermedios que permiten replicar el proceso; en Apriori se documentaron los umbrales de soporte y confianza y se almacenaron las reglas y sus visualizaciones, mientras que en clustering se aplicaron métricas como Silhouette Score junto con la conservación de resultados numéricos y gráficos de evaluación.

A nivel organizacional, GitHub fue utilizado como sistema de control de versiones para respaldar el avance, registrar modificaciones y documentar decisiones técnicas mediante archivos .md, reforzando la trazabilidad. Además, la integración de

resultados en un dashboard interactivo actuó como evidencia visual del funcionamiento conjunto de los modelos, mostrando su potencial aplicación en entornos reales de apoyo a decisiones.

En conjunto, estos mecanismos aseguran que cada fase del proyecto sea verificable, reproducible y alineada con los objetivos metodológicos y de calidad establecidos.

5- Plan de gestión/proyecto (calidad y testing)

5.1 Gestión de la calidad

La calidad en este proyecto se aborda tanto desde la confiabilidad del proceso analítico como desde la pertinencia de los entregables.

- Calidad del proceso: Se garantizó mediante la trazabilidad completa del flujo de trabajo, la aplicación de técnicas de limpieza sistemática de datos, el uso de librerías científicas consolidadas en Python (como pandas, mlxtend, scikit-learn y plotly) y la documentación cuidadosa de cada etapa mediante comentarios de código y estructuras de carpetas organizadas.

- Calidad del entregable final: Se controló a través de revisiones iterativas del informe escrito, validación cruzada entre resultados y objetivos, y el seguimiento del cumplimiento del plan de trabajo. Además, el desarrollo de visualizaciones y resúmenes tabulares reforzó la interpretación coherente de resultados, fortaleciendo la utilidad práctica del análisis.

5.2 Plan de riesgos

Riesgo identificado	Impacto	Probabilidad	Estrategia de mitigación
Baja calidad del dataset seleccionado	Alto	Media	Evaluar múltiples datasets, realizar limpieza exhaustiva y aplicar validaciones previas.
Dificultades en la implementación de algoritmos	Medio	Alta	Consultar documentación técnica, foros especializados y realizar pruebas incrementales.
Exceso de carga académica externa al proyecto	Alto	Alta	Establecer un cronograma realista con bloques de trabajo semanales.
Resultados no interpretables o poco relevantes	Medio	Media	Ajustar parámetros, revisar supuestos y redefinir objetivos analíticos si es necesario.

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

5.3 Plan de testing

La validación técnica de los análisis se realizó a través de pruebas específicas para cada técnica aplicada, centradas en la robustez del modelo, la claridad de los resultados y su valor comercial potencial.

5.4 Testing para reglas de asociación (Apriori)

- Validación del soporte y confianza mínimo definido (≥ 5% y ≥ 70%, respectivamente).
- Revisión manual de reglas para excluir asociaciones triviales o redundantes.

- Exportación de las reglas relevantes a archivos .csv y su visualización mediante gráficos de red y matrices de calor para evaluar su significado práctico.

5.5 Testing para clustering (K-means):

- Evaluación del número óptimo de clústeres mediante el método del codo y Silhouette Score.
- Validación visual a través de reducción de dimensionalidad con PCA y gráficas comparativas por variable.
- Revisión del sentido comercial de cada grupo, considerando patrones de compra, departamentos preferidos y comportamiento agregado.

Este enfoque de testing permitió verificar no solo la ejecución técnica correcta, sino también la capacidad explicativa de los resultados, asegurando su relevancia para la estrategia comercial de un supermercado online como Lider.cl.

6- Plan de riesgos

En este proyecto se aplican los principios del marco PMBOK para estructurar de manera coherente la gestión del trabajo, integrando alcance, riesgos, calidad y comunicaciones con el fin de orientar el desarrollo con claridad, reducir incertidumbres y asegurar que cada actividad contribuya a los objetivos definidos. El alcance se centra en diseñar y validar un sistema analítico capaz de identificar patrones de compra y segmentar clientes a partir de datos simulados, sin integrar plataformas reales ni tratar información privada. Los entregables contemplan scripts reproducibles, visualizaciones interpretables, archivos de resultados y un dashboard demostrativo, todo documentado en un informe técnico que expone la metodología y los hallazgos principales.

Los riesgos se identificaron desde el inicio y cuentan con estrategias tanto preventivas como reactivas; su revisión es continua en el cronograma y las decisiones relevantes se registran en el repositorio, lo que permite respuestas ágiles ante contingencias y garantiza la disponibilidad de entregables mínimos viables en cada fase. La gestión de la calidad combina la validación técnica —limpieza de datos, trazabilidad del flujo de trabajo y evaluación de modelos mediante métricas como soporte, confianza y Silhouette Score— con la consistencia del producto final, velando por la claridad metodológica del informe, las visualizaciones y el dashboard.

Finalmente, el plan de comunicaciones articula la coordinación interna mediante Visual Studio Code y control de versiones con Git, junto con el cumplimiento de los formatos académicos y las instancias de presentación y retroalimentación. En conjunto, estos elementos sostienen una ejecución controlada, metodológicamente consistente y orientada a obtener resultados útiles, reproducibles y acordes con estándares profesionales.

7- Cronograma del proyecto

El cronograma del proyecto fue diseñado en base a la metodología CRISP-DM, articulada mediante una estructura iterativa de sprints semanales inspirada en los principios de gestión ágil. Este enfoque permite avanzar de forma progresiva en cada fase del análisis de datos, con espacios definidos para revisión y ajustes.

El desarrollo se proyectó en un total de 8 semanas, adecuándose tanto a la carga académica como a los plazos establecidos por la asignatura. Cada semana se estructuró como un sprint con metas específicas, tareas técnicas delimitadas y cierre con validación de avances. A continuación, se presenta el plan de trabajo general:

Semana	Fase del proyecto	Actividades clave
1	**Definición del problema**	Redacción inicial del planteamiento, formulación de objetivos, revisión bibliográfica.
2	**Comprensión de datos**	Selección del dataset, análisis exploratorio preliminar, identificación de variables.
3	**Preparación de datos**	Limpieza de registros, transformación de variables, codificación de campos.
4	**Modelado – Asociación**	Implementación de Apriori, extracción de reglas, validación por soporte y confianza.
5	**Modelado – Clustering**	Aplicación de K-means, evaluación por Silhouette Score, visualización de resultados.
6	**Evaluación de resultados**	Interpretación cruzada de clústeres y reglas, extracción de insights comerciales.
7	**Prototipo y visualización**	Construcción del dashboard, refinamiento de gráficos, diseño conceptual de interfaz.
8	**Cierre y entrega**	Redacción final del informe, revisión del repositorio, envío de productos y anexos.

Tabla 3: Cronograma del proyecto.

Durante el desarrollo, se utilizó Visual Studio Code como entorno principal, junto con el control de versiones en GitHub, lo que permitió mantener una trazabilidad efectiva del avance. Se complementó con checklists semanales para monitorear el cumplimiento de tareas y facilitar la detección temprana de desvíos.

Esta planificación estructurada aseguró un avance constante, incorporando espacios de validación, corrección y documentación que garantizaron la calidad del entregable final y su coherencia con los objetivos propuestos.

8- Prototipo

Con el objetivo de validar la viabilidad técnica y visual de la solución propuesta, se construyó un prototipo funcional que permite ejecutar y visualizar parte de los resultados generados por el sistema analítico. El prototipo se desarrolló en un entorno local, utilizando Python como lenguaje principal, y se apoya en visualizaciones generadas con librerías como matplotlib, seaborn, networkx y plotly.

Este prototipo no busca reemplazar una aplicación comercial, sino ofrecer una instancia de exploración de resultados en tiempo real, accesible para usuarios con conocimientos técnicos intermedios. La estructura modular del proyecto permite reproducir fácilmente los pasos de análisis y visualizar los resultados en distintas etapas del proceso. A continuación, se presenta una captura del prototipo ejecutado en local:

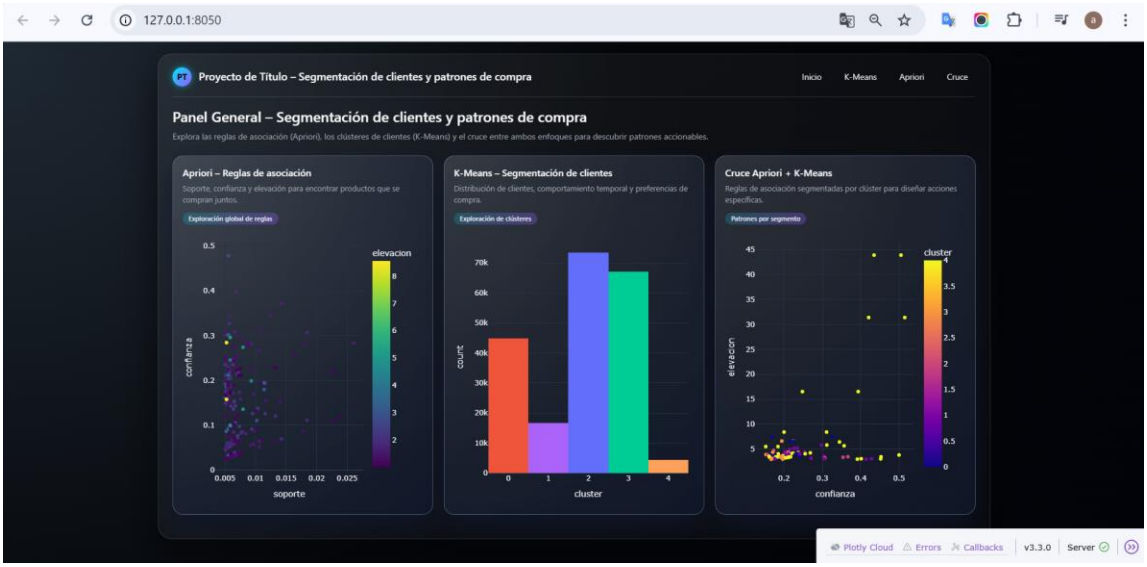

Figura 3: Prototipo ejecutado en local.

CAPÍTULO IV: DISCUSIÓN DE RESULTADOS

1- Gestión de proyecto

Durante el desarrollo del proyecto, se aplicó de manera estructurada y consistente una metodología híbrida, que combinó dos enfoques complementarios:

- CRISP-DM (para el desarrollo del producto analítico).
- Scrum + PMBOK (para la planificación, ejecución y control del proyecto).

Ambos marcos metodológicos se vincularon estrechamente al cronograma de trabajo original, permitiendo una ejecución escalonada, con validaciones progresivas en cada iteración.

1.1 Aplicación de CRISP-DM

1. Comprensión del negocio: Se definieron los problemas centrales del contexto de supermercados online, y se formularon los objetivos analíticos. Esta etapa quedó reflejada en la fundamentación, objetivos y el planteamiento del problema abordado.
2. Comprensión de los datos: Se evaluaron distintos datasets públicos, seleccionando aquel más representativo para el objetivo del proyecto. Se exploraron dimensiones como fechas, productos, precios y usuarios, evaluando su potencial analítico y calidad estructural.
3. Preparación de los datos: Se realizó limpieza, transformación y codificación de las variables. Se estandarizaron campos, se depuraron registros incompletos, y se generaron nuevas columnas derivadas, como rangos horarios, totales por cliente, proporciones de productos por categoría, entre otras.
4. Modelado: Se aplicaron dos técnicas principales:

- Reglas de asociación (Apriori): para descubrir productos que se compran juntos.
- Clustering (K-means): para segmentar clientes en grupos con comportamientos similares.

Adicionalmente, se implementó un cruce exploratorio entre ambos modelos, con el fin de identificar asociaciones diferenciadas por clúster.

5. Evaluación: Los resultados fueron validados mediante métricas clave:
   - Soporte y confianza para las reglas Apriori.
   - Silhouette Score y análisis PCA para los clústeres.
   - Relevancia comercial de las asociaciones.

Además, se evaluó la coherencia semántica de las agrupaciones de clientes y la utilidad potencial de los hallazgos para la toma de decisiones estratégicas.

6. Despliegue (simulado): Los resultados se integraron en un dashboard exploratorio construido con Dash, y se documentaron en informes técnicos y visuales para su interpretación por usuarios no técnicos.

1.2 Aplicación de gestión ágil (Scrum + PMBOK)

El desarrollo fue organizado en 8 iteraciones semanales que siguieron la lógica de Sprints, cada uno con entregables definidos, revisión al cierre y ajuste de prioridades según resultados intermedios. Estas iteraciones están alineadas con las fases CRISP-DM y permitieron un avance continuo y controlado.

El uso de elementos del PMBOK como soporte permitió estructurar los siguientes aspectos:

- Plan de calidad: Validaciones técnicas y documentales.
- Plan de cronograma: Control semanal de tareas y fechas clave.
- Gestión de alcance: Mantención del enfoque analítico frente a nuevas ideas surgidas durante el desarrollo.
- Gestión de riesgos: Anticipación de obstáculos y decisiones técnicas informadas.

Se mantuvo un registro de decisiones en GitHub y un seguimiento semanal de hitos, lo que permitió gestionar adecuadamente los tiempos y mitigar desviaciones en etapas críticas.

2- Diseño de componentes funcionales

El diseño del sistema desarrollado responde a la lógica de un proyecto centrado en análisis de datos y construcción de conocimiento aplicado. A diferencia de un software transaccional o productivo, esta solución se estructura en torno a flujos analíticos, con un único perfil de usuario (analista de datos) que ejecuta las distintas etapas según el modelo CRISP-DM.

2.1 Diagrama de casos de uso

El sistema contempla un conjunto acotado de acciones realizadas por el analista. Entre ellas se incluyen la carga y preparación de datos, la ejecución de algoritmos (Apriori y K-means), la interpretación de resultados y la generación de productos visuales (gráficos, tablas, reportes). Estas acciones están representadas en el siguiente diagrama de casos de uso:

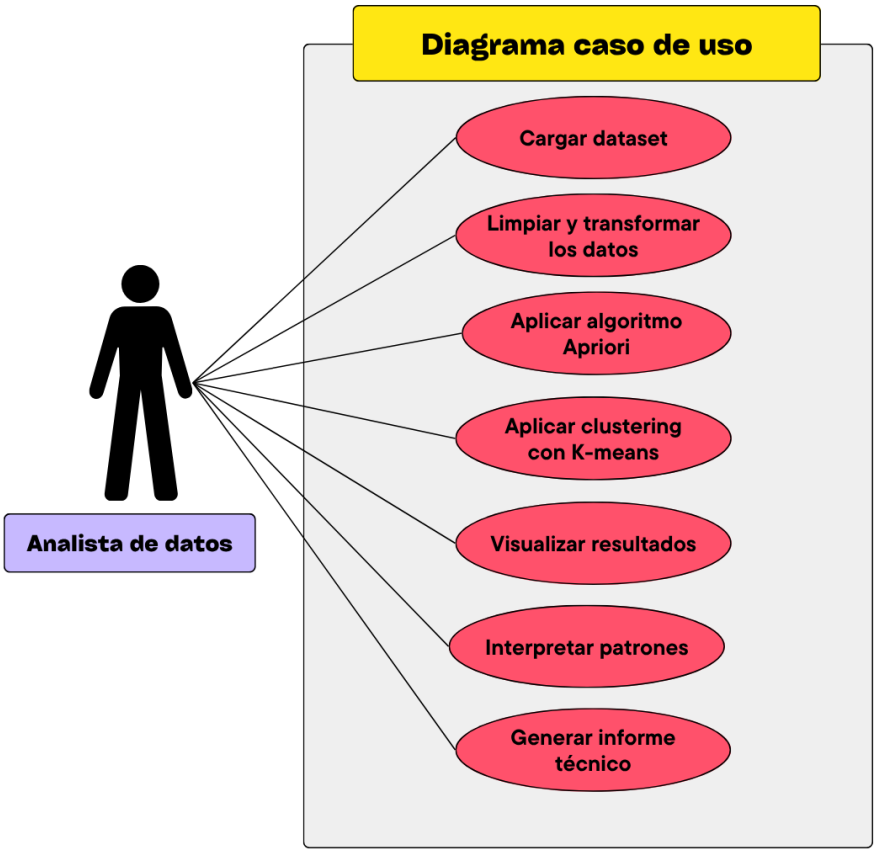

Figura 4: Diagrama de caso de uso.

2.2 Vistas del sistema según el modelo 4+1

Dado que la solución no es una aplicación desplegada, sino un entorno analítico de ejecución local se opta por representar la arquitectura desde una perspectiva académica utilizando el modelo de vistas 4+1 adaptado.

2.3 Vista lógica

Refleja los componentes principales del sistema, organizados en bloques funcionales:

- Módulo de carga y limpieza de datos.

- Módulo de reglas de asociación (Apriori).
- Módulo de clustering (K-means).
- Módulo de análisis cruzado y generación de visualizaciones.

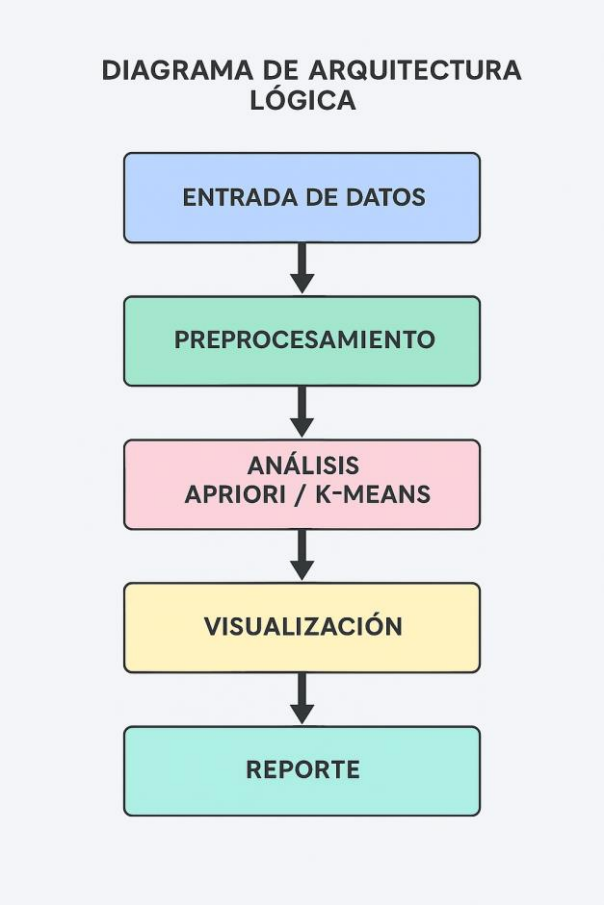

Figura 5: de arquitectura lógica.

2.4 Vista de desarrollo

La estructura del proyecto sigue un enfoque modular, organizado en carpetas según el tipo de análisis (Apriori, K-means, cruces) y separados por función (scripts/, output/, data/). Esta organización permite una gestión eficiente del código, los datos y los resultados visuales generados.

A continuación, se incluye una captura del entorno de trabajo que refleja esta estructura de carpetas implementada en Visual Studio Code:

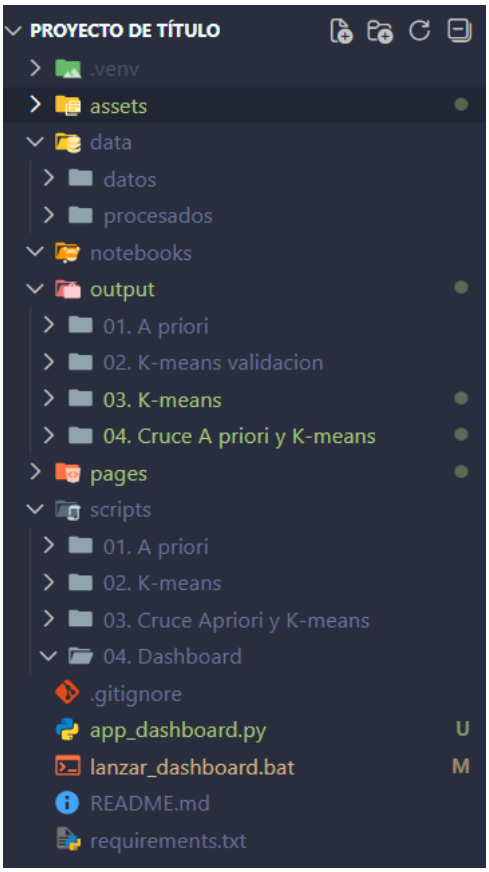

Figura 6: Estructura modular del proyecto en VS Code.

2.5 Vista de procesos

El flujo lógico de ejecución del sistema se estructura en etapas secuenciales, siguiendo el modelo CRISP-DM:

1. Carga y revisión del conjunto de datos (archivos CSV simulados de compras).
2. Limpieza y transformación de datos (preprocesamiento).
3. Aplicación de algoritmos: reglas de asociación (Apriori) y clustering (K-means).
4. Generación de visualizaciones (gráficos PNG por técnica y por clúster).

5. Implementación de un dashboard exploratorio (Dash + Plotly).
6. Interpretación de resultados y análisis cualitativo.

Este proceso se encuentra automatizado en notebooks y scripts Python, con salidas organizadas para facilitar su análisis por etapas.

2.6 Vista física

El sistema fue desarrollado y ejecutado localmente, sin necesidad de infraestructura en la nube. La plataforma empleada para el desarrollo y ejecución fue:

- Visual Studio Code con terminal Bash (Git Bash).
- Entorno virtual Python 3.12.6, activado manualmente mediante script.
- Servidor local levantado en localhost:8050 mediante Dash y Flask.
- Ejecución multiplataforma compatible con entornos Windows.

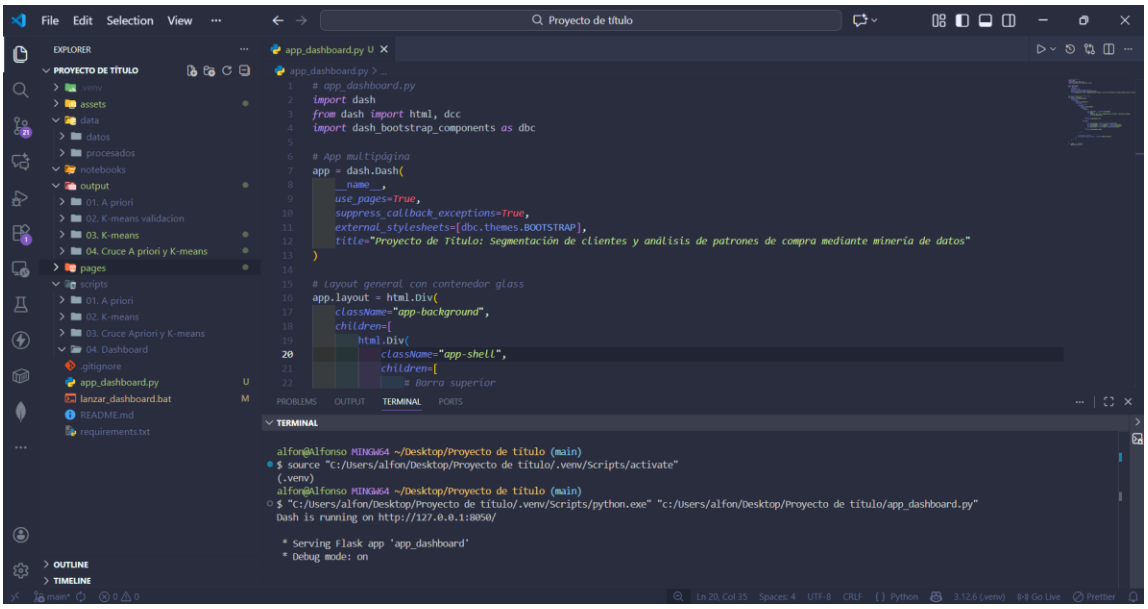

Figura 7: Ejecución local del dashboard desde entorno virtual en VSC.

3- Matriz de trazabilidad de requerimientos

ID	Requerimiento	Sprint / Etapa CRISP-DM	Componente de diseño asociado	Implementación
RF01	Identificar patrones de compra mediante reglas de asociación	Modelado (Apriori)	Caso de uso: Generar reglas frecuentes	scripts/01. A priori, output/01.*
RF02	Segmentar clientes con base en variables de comportamiento	Modelado (K-means)	Caso de uso: Agrupar clientes por clúster	scripts/02. K-means, output/03.*
RF03	Explorar relaciones entre reglas y segmentos	Cruce Apriori-Kmeans	Caso de uso: Filtrar reglas por clúster	scripts/03. Cruce*, output/04.*
RF04	Visualizar métricas y patrones en gráficos interpretables	Despliegue (visualización)	Arquitectura lógica y vistas	output/*.png, dashboard Plotly Dash
RF05	Organizar resultados en carpetas separadas según técnica aplicada	Desarrollo	Vista de desarrollo	Estructura carpetas scripts/output
RF06	Permitir ejecución local desde terminal con entorno controlado	Despliegue físico	Vista física	lanzar_dashboard.bat, terminal Bash
RNF01	Utilizar solo software y librerías open source	General	Restricción tecnológica	Python + requirements.txt
RNF02	Garantizar trazabilidad del flujo de datos y código	Todo el ciclo	Revisión y documentación	README.md, comentarios
RNF03	Entorno reproducible y versionable	Configuración inicial	Control de versiones y ambiente	.gitignore, .venv, estructura

Figura 8: Matriz de trazabilidad de requerimientos.

4- Diseño y construcción del producto de software

4.1 Diseño general del sistema

El diseño del sistema desarrollado sigue una estructura analítica centrada en el procesamiento de datos y la generación de conocimientos accionables a partir de técnicas de minería de datos. Considerando que el objetivo principal es identificar patrones de comportamiento y segmentar clientes para un supermercado online, el sistema se compone de componentes que operan de forma secuencial, pero desacoplada, desde la lectura de datos hasta la visualización final.

El sistema fue concebido para ejecutarse en entorno local, utilizando principalmente VS Code con scripts en Python. El desarrollo se apoya en la metodología CRISP-DM, por lo que la estructura modular del proyecto refleja cada una de las fases: comprensión, preparación, modelado, evaluación y presentación de resultados.
Los bloques funcionales del sistema incluyen:

- Módulos de asociación (Apriori): generación de reglas, visualización y análisis de patrones de coocurrencia.

- Módulos de clustering (K-means): validación del número óptimo de grupos, creación de clústeres, análisis de comportamiento por segmento.
- Cruce Apriori-K-means: integración de ambos enfoques para enriquecer el análisis según la segmentación.
- Dashboard final: interfaz visual interactiva para resumir resultados clave.

Además, el sistema fue diseñado bajo un enfoque modular. Cada componente del análisis se encuentra organizado en carpetas independientes según su función: scripts de Apriori, K-means, cruce de técnicas y dashboard, lo que permite mantener el código limpio, comprensible y escalable. Las salidas gráficas se almacenan en carpetas específicas por técnica, mientras que los resultados intermedios y finales (archivos .csv) se ubican en una ruta dedicada a datos procesados. Esta organización refleja una arquitectura lógica orientada a tareas, favorece la reproducibilidad y permite que el sistema sea fácilmente extendido o adaptado a nuevos análisis.

En conjunto, este diseño garantiza la trazabilidad del flujo analítico, facilita el control de calidad en cada etapa del desarrollo y permite la entrega de un sistema profesional, replicable y comprensible para futuros análisis.

4.2 2 Diseño de la base de datos

El sistema no utiliza una base de datos tradicional, sino una estructura modular basada en archivos CSV organizados en tres niveles: datos fuente, datos procesados y salidas visuales. Los archivos en data/datos/ representan tablas base simuladas del supermercado, mientras que data/procesados/ contiene resultados intermedios como clientes clusterizados o reglas de asociación. Esta estructura permite trazabilidad, reutilización y replicación del flujo de trabajo sin necesidad de un gestor de base de datos. Las relaciones se establecen por claves implícitas

(user_id, order_id, product_id), cumpliendo el rol funcional de una base de datos relacional adaptada al entorno analítico.

4.3 Diseño de la plataforma de operación

El sistema fue desarrollado y ejecutado en un entorno local, usando exclusivamente Visual Studio Code como plataforma de trabajo principal. La operación del sistema se distribuye de la siguiente forma:

- Procesamiento de datos y ejecución de scripts: Todo el procesamiento (Apriori, K-means, visualizaciones) se ejecuta en Python a través de scripts modulares organizados en la carpeta scripts/, agrupados según funcionalidad.
- Visualización y análisis: Los resultados generados se almacenan como archivos CSV y gráficos .png, y son interpretados desde notebooks de Jupyter o directamente desde scripts.
- Dashboard interactivo: Se diseñó un panel visual de exploración con Dash (scripts/04. Dashboard/), el cual se lanza localmente mediante un script .bat, sin necesidad de despliegue web.
- Control de versiones y respaldo: Todo el proyecto fue versionado en Git y respaldado en GitHub, asegurando trazabilidad del desarrollo, documentación técnica y disponibilidad del código.

No se utilizaron servidores externos ni bases de datos desplegadas. La plataforma está pensada para reproducirse íntegramente en entornos de escritorio con Python y VSC configurados.

4.4 Resultados gráficos generados por el sistema analítico

4.4.1 Reglas de asociación con A priori

Durante la aplicación del algoritmo Apriori, se identificaron múltiples patrones de compra a partir del dataset transaccional. A continuación, se presentan los principales gráficos generados durante esta fase del sistema analítico:

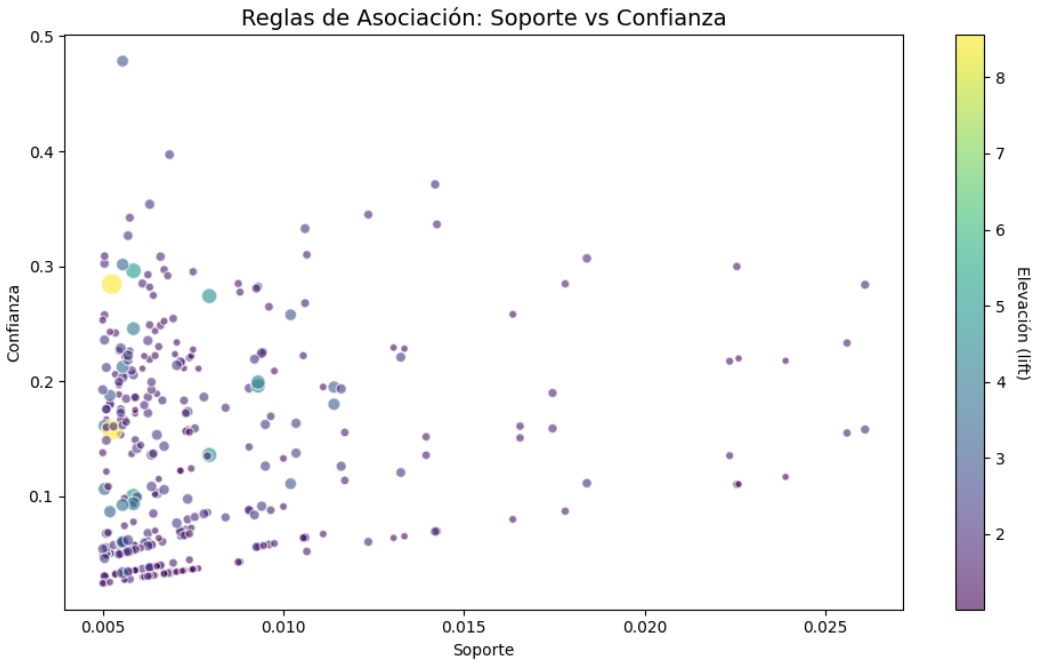

Figura 9: Reglas de asociación generadas por A priori.

Este gráfico muestra el total de reglas encontradas ordenadas por soporte, confianza y lift. Permite visualizar cuáles combinaciones de productos superaron los umbrales establecidos y podrían tener valor para una estrategia de recomendación o promoción.

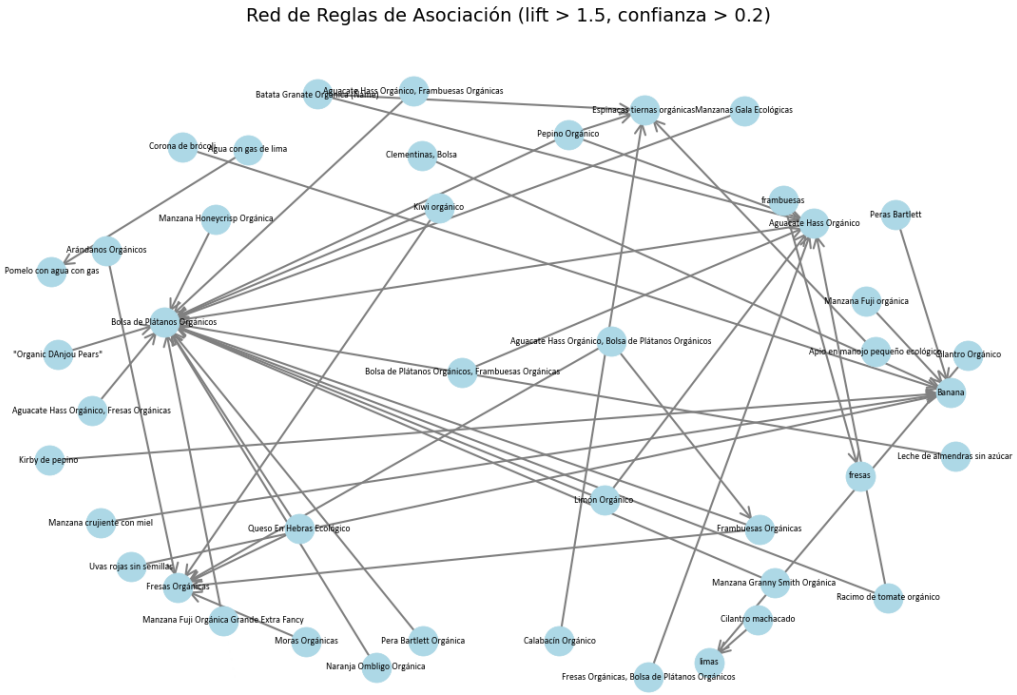

Figura 10: Red amplia de reglas de asociación.

Se visualiza la totalidad de reglas válidas como una red de nodos. Cada nodo representa un producto y las conexiones indican asociaciones frecuentes. Este tipo de visualización resulta útil para identificar comunidades de productos relacionados.

Figura 11: Red reducida de reglas (reglas más relevantes).

A diferencia del gráfico anterior, aquí se presenta una selección filtrada de las reglas con mayor confianza y lift. Se reduce la complejidad visual para facilitar la lectura e interpretación de los patrones más fuertes.

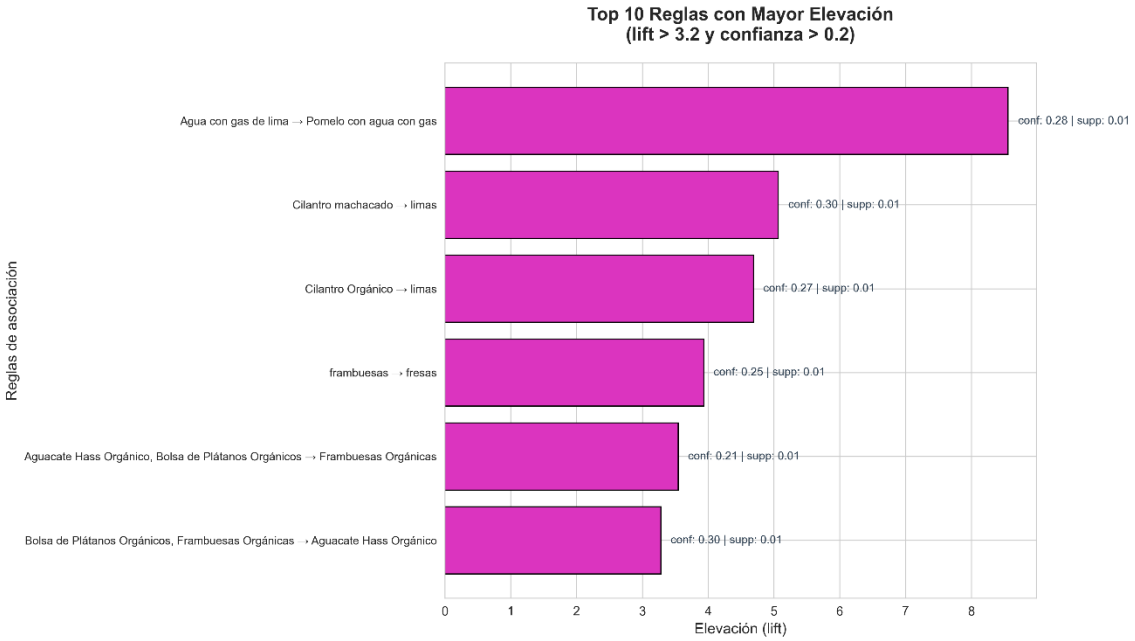

Figura 12: Distribución de reglas por producto.

Este gráfico de barras muestra la cantidad de veces que cada producto aparece en las reglas generadas. Permite detectar productos con fuerte protagonismo dentro de las asociaciones frecuentes, lo que puede guiar decisiones de posicionamiento.

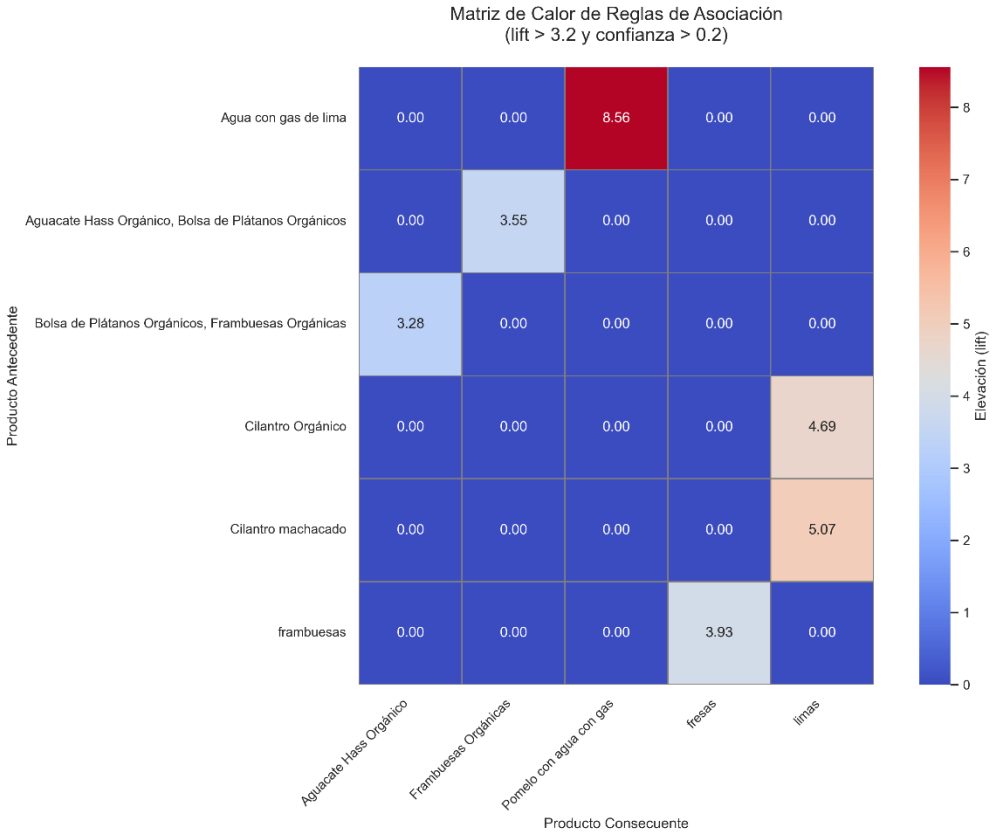

Figura 13: Matriz de calor de productos asociados.

Se presenta una matriz de calor que indica la frecuencia de coocurrencia entre productos. Cuanto más intenso el color, mayor la frecuencia de aparición conjunta. Esta visualización complementa la red al proporcionar una visión más cuantitativa.

La serie de gráficos generados en esta etapa permitió validar visualmente la consistencia del modelo de reglas. Se logró identificar productos centrales en las asociaciones, relaciones fuertes entre ciertos pares y comunidades de productos que podrían sugerirse juntos. La combinación entre visualización de red, matriz y distribución por producto aportó una comprensión integral de los patrones hallados por el sistema.

4.4.2 Clustering con K-means

Luego de aplicar el algoritmo K-means sobre el conjunto de variables seleccionadas, se obtuvieron cinco grupos de clientes con comportamientos diferenciados. A continuación, se presentan los principales gráficos generados para explorar y caracterizar dichos clústeres:

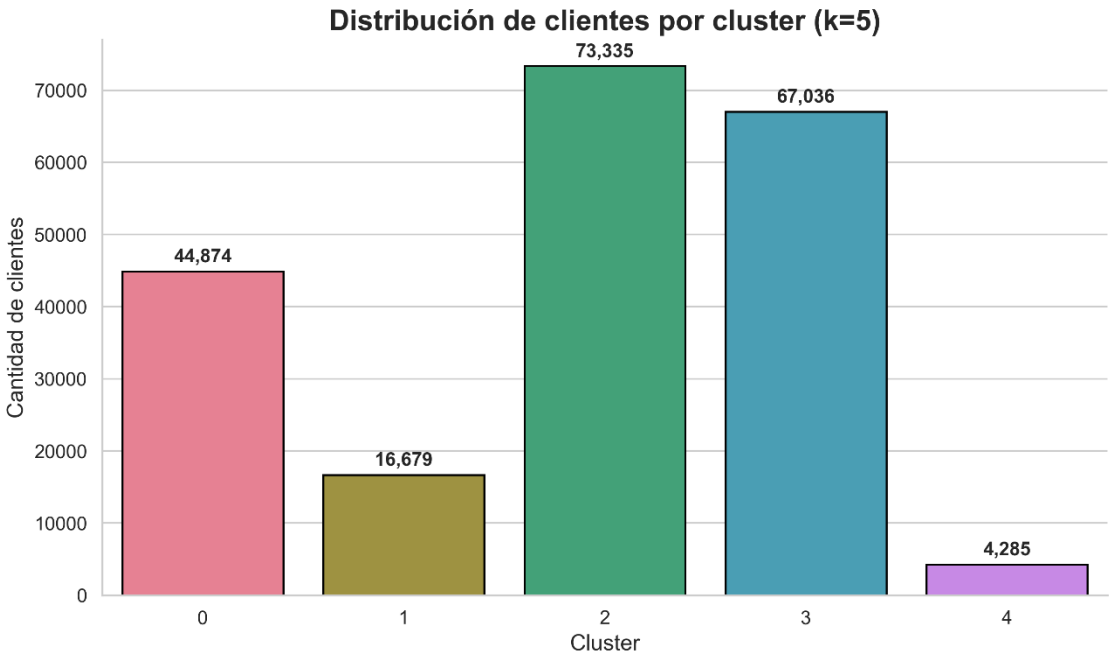

Figura 14: Distribución de clientes por clúster.

Este gráfico muestra el número de clientes que pertenecen a cada clúster, permitiendo observar la proporción relativa de cada grupo. Es útil para detectar si la segmentación está equilibrada o si existen clústeres dominantes.

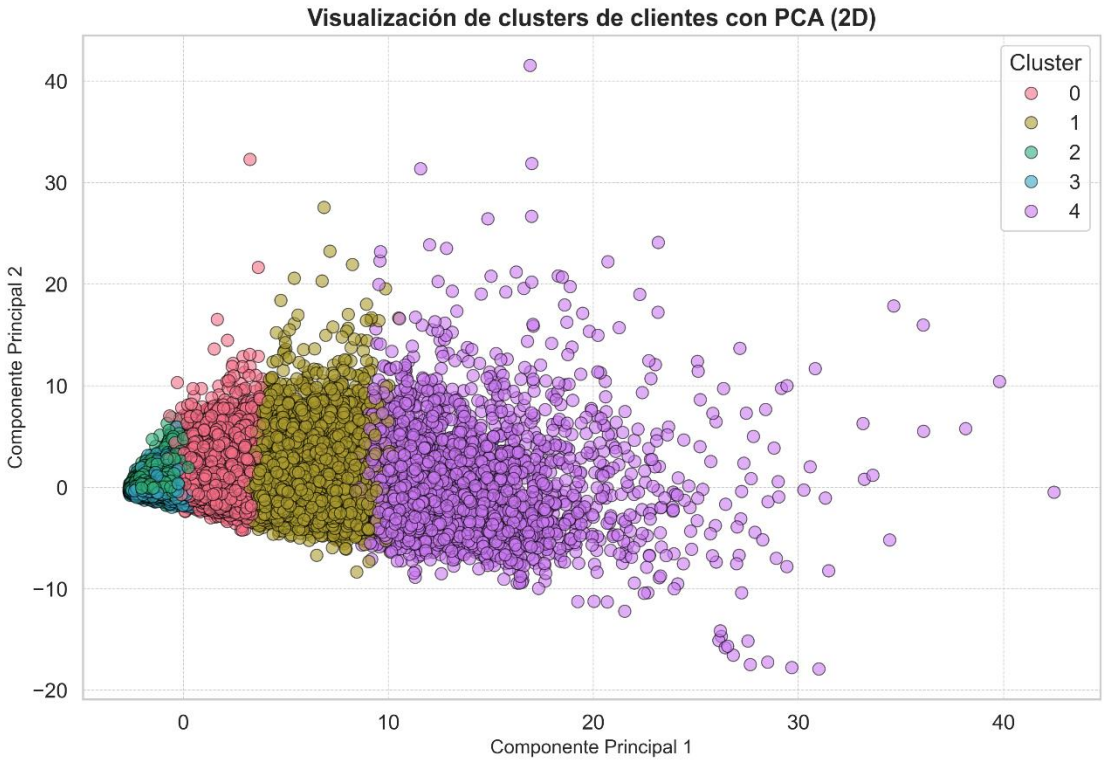

Figura 15: Visualización de clústeres de cliente con PCA.

Utilizando Análisis de Componentes Principales (PCA), se redujo la dimensionalidad del dataset para representar los clústeres en un plano 2D. Este gráfico permite validar visualmente la separación y cohesión interna de los grupos generados por K-means.

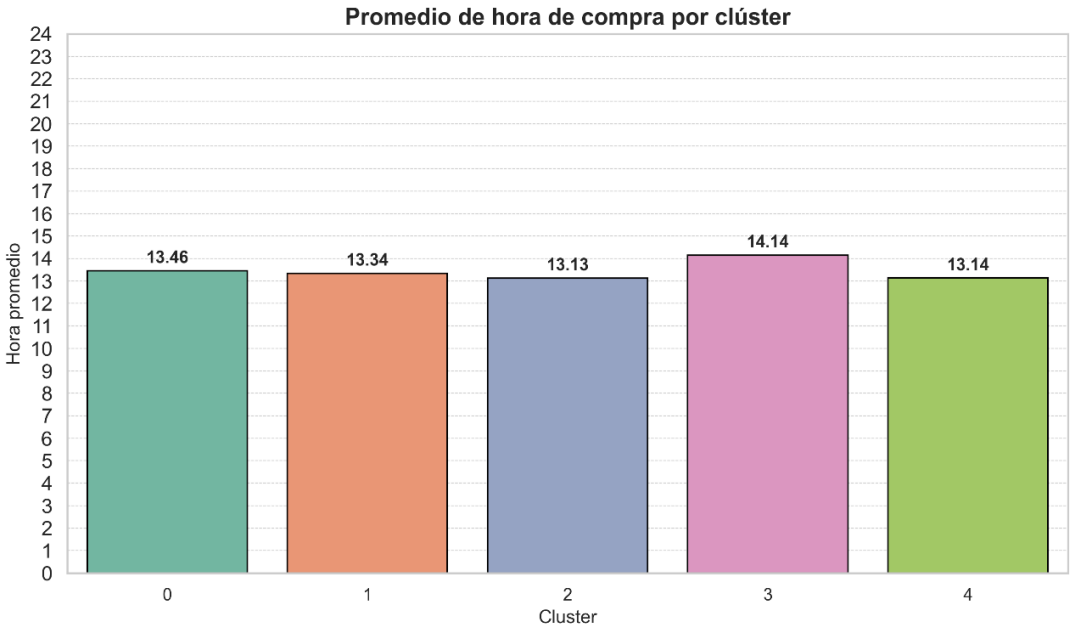

Figura 16: Promedio de hora de compra por clúster.

Se observa la hora promedio en la que los clientes de cada clúster realizan sus pedidos. Esto permite detectar patrones temporales y planificar acciones comerciales según franjas horarias preferentes por grupo.

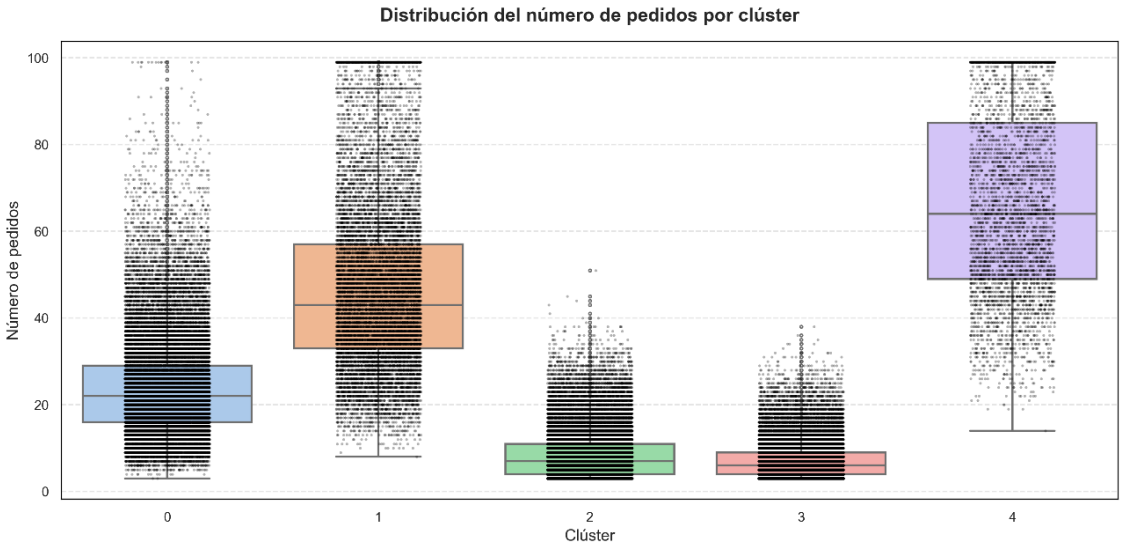

Figura 17: Número de pedidos por clúster.

Muestra la cantidad promedio de pedidos realizados por los clientes de cada clúster, lo que da cuenta de su nivel de actividad o recurrencia dentro del sistema de compras.

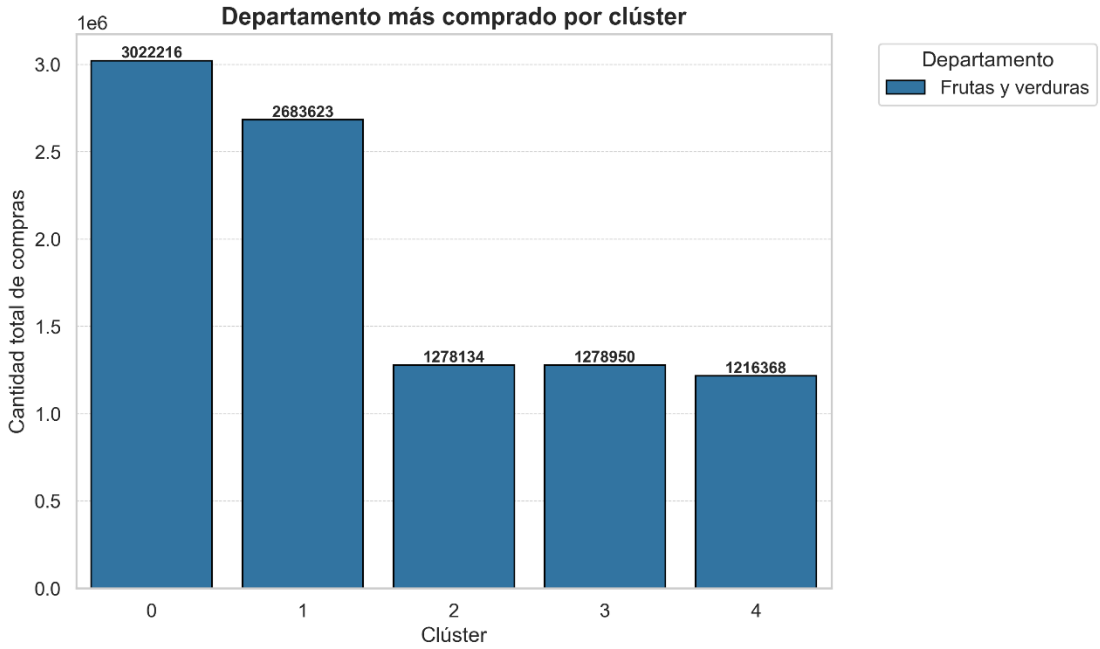

Figura 18: Departamento más comprado por clúster.

Identifica, para cada clúster, cuál es el departamento (categoría de producto) con mayor frecuencia de compra. Aporta una visión directa de las preferencias dominantes por segmento.

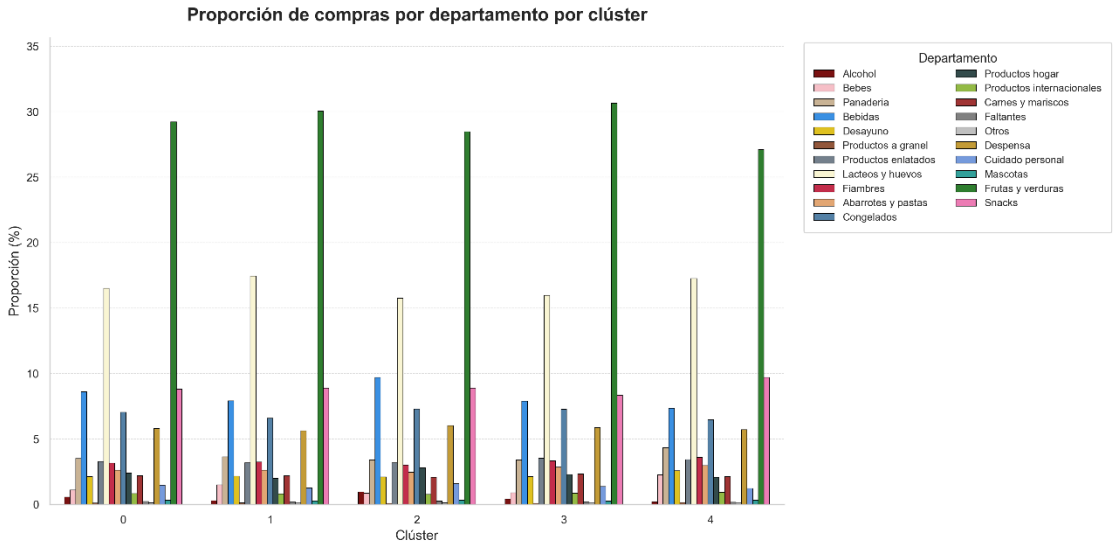

Figura 19: Proporción de compras por departamento por clúster.

A diferencia del anterior, este gráfico representa la proporción completa de compras realizadas por cada clúster distribuidas por departamento, permitiendo comparar preferencias relativas entre grupos.

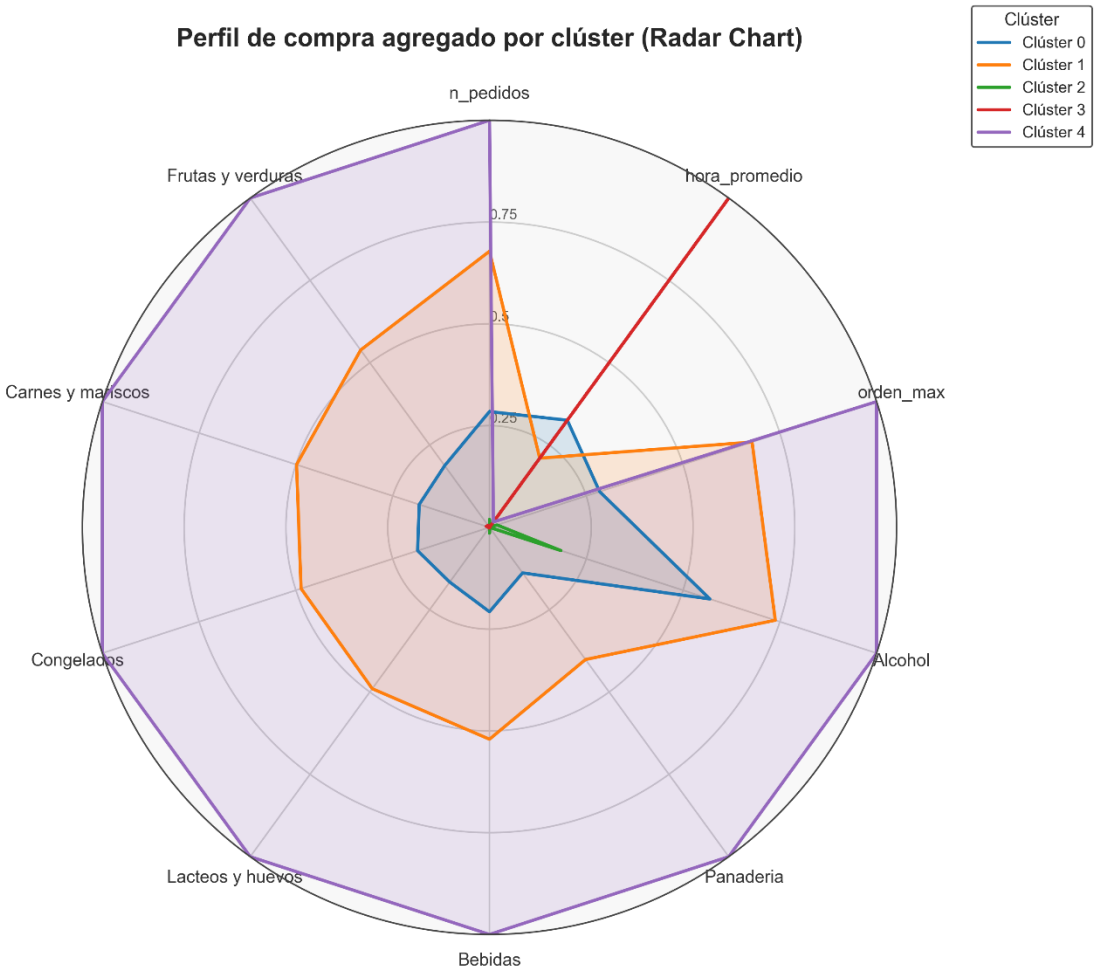

Figura 20: Perfil agregado. Radar Chart por clúster.

Integra múltiples dimensiones de comportamiento (hora de compra, volumen, preferencia, gasto) en una sola visualización tipo radar. Facilita la comparación directa de perfiles globales de cada clúster, destacando sus fortalezas y diferencias.

La segmentación por K-means permitió construir perfiles de clientes significativamente diferenciados en cuanto a comportamiento, preferencias y volumen de compra. Las visualizaciones evidencian que los clústeres no solo

difieren en tamaño, sino también en patrones horarios y categorías de consumo. Estas diferencias abren posibilidades de personalización en recomendaciones, promociones y gestión operativa. El uso combinado de técnicas estadísticas y gráficas asegura que la segmentación no sea solo numérica, sino también interpretable

4.4.3 Cruce Apriori – K-means

Esta fase integra los resultados del clustering con la minería de reglas de asociación. Para cada clúster identificado con K-means, se extrajo un subconjunto de transacciones y se aplicó nuevamente el algoritmo Apriori. De este modo, se obtuvieron reglas específicas por segmento de cliente, lo que permite recomendaciones personalizadas.

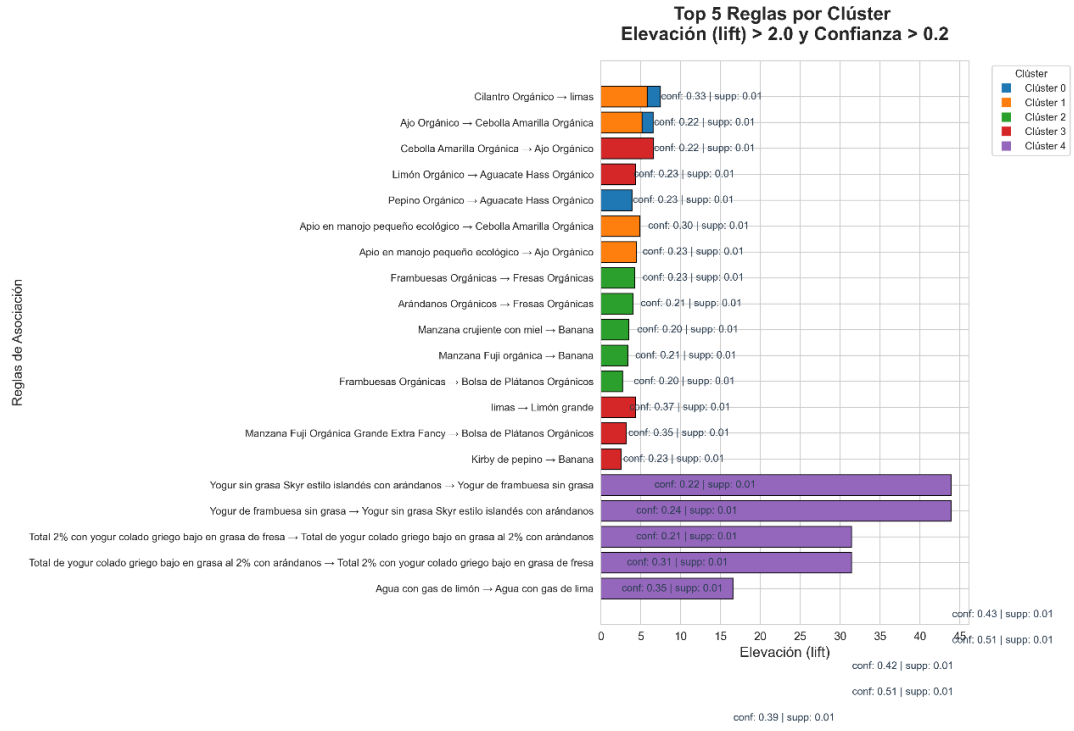

Figura 21: Top reglas por clúster.

Este gráfico muestra, en una tabla visual, las principales reglas de asociación generadas para cada uno de los cinco clústeres, ordenadas por soporte y confianza. Sirve como resumen comparativo para identificar la naturaleza de las asociaciones más relevantes por perfil de cliente.

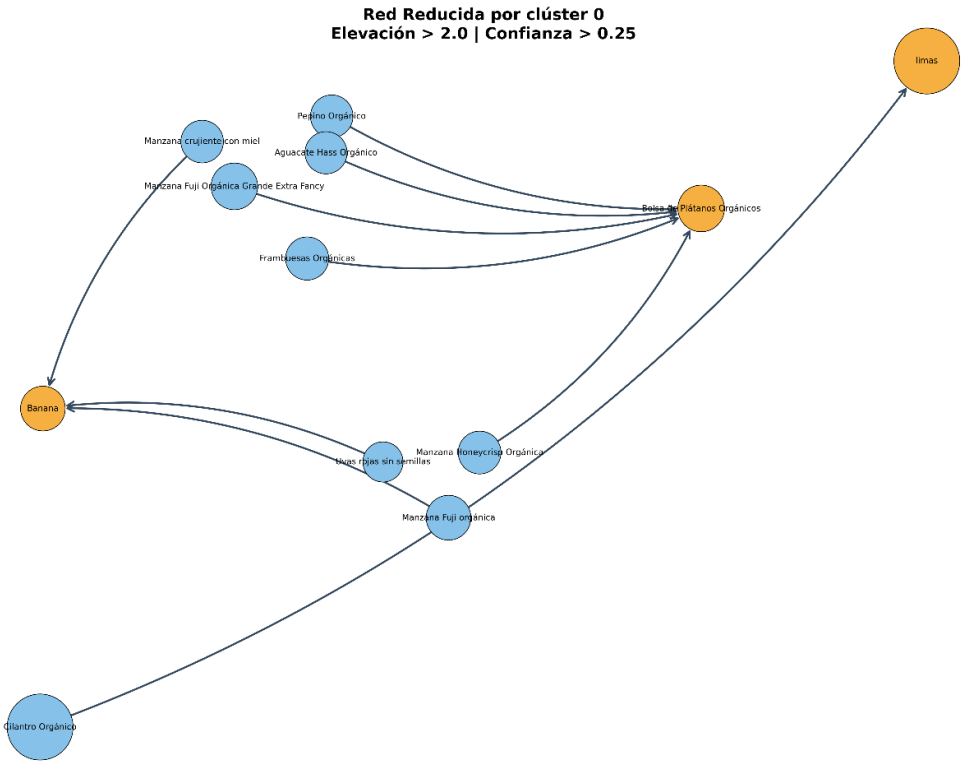

Figura 22: Red de reglas por clúster 0.

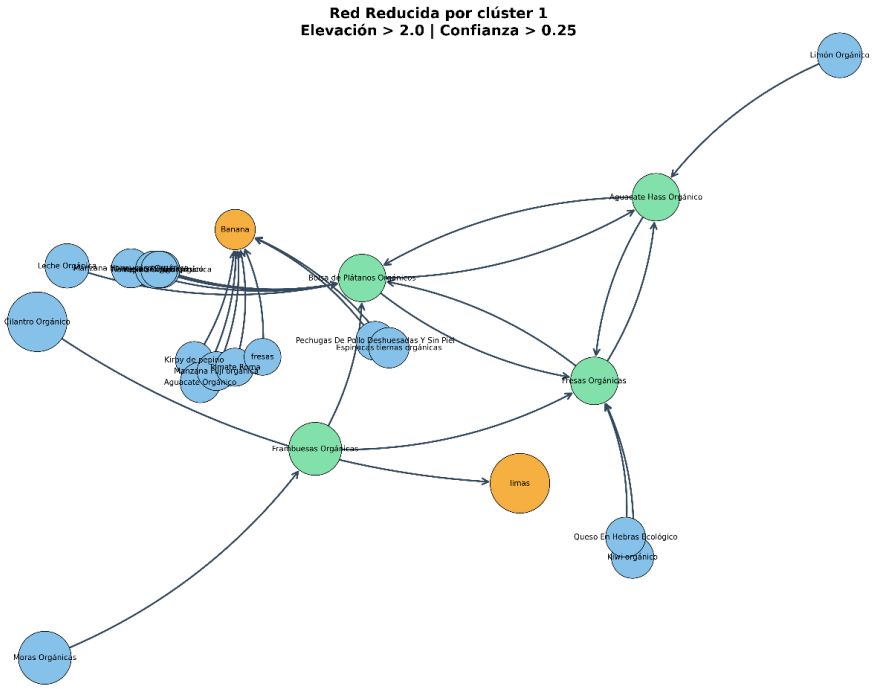

Figura 23: Red de reglas por clúster 1.

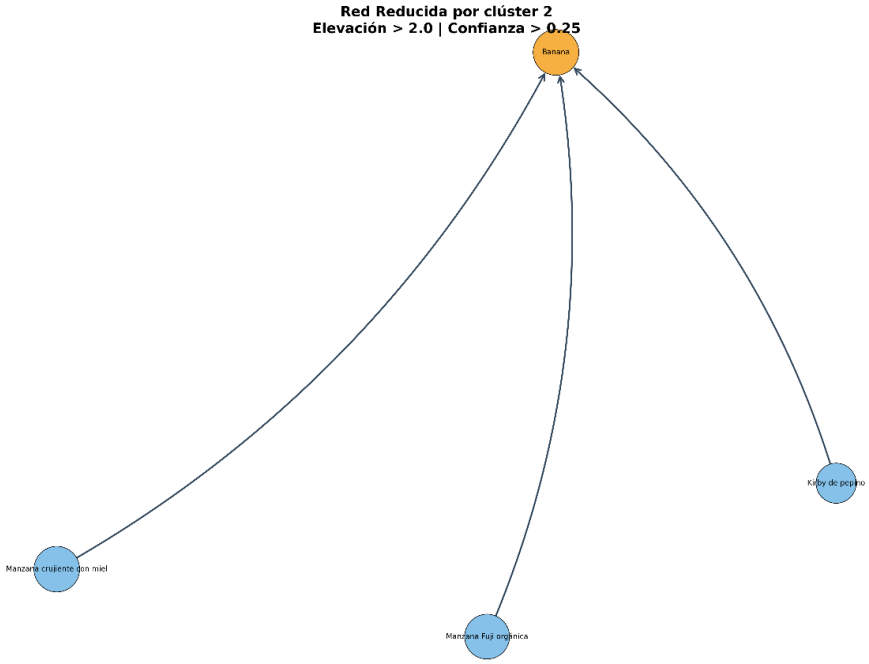

Figura 24: Red de reglas por clúster 2.

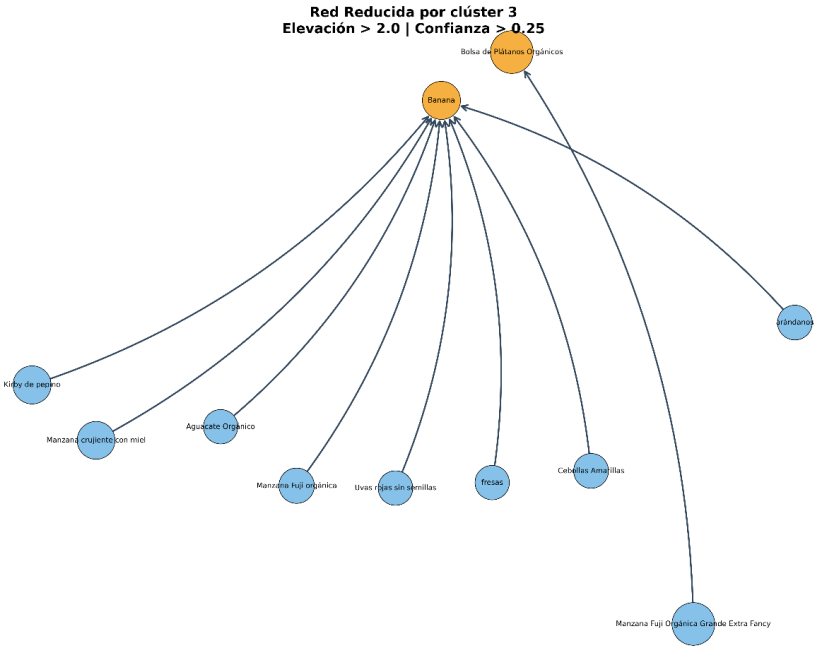

Figura 25: Red de reglas por clúster 3.

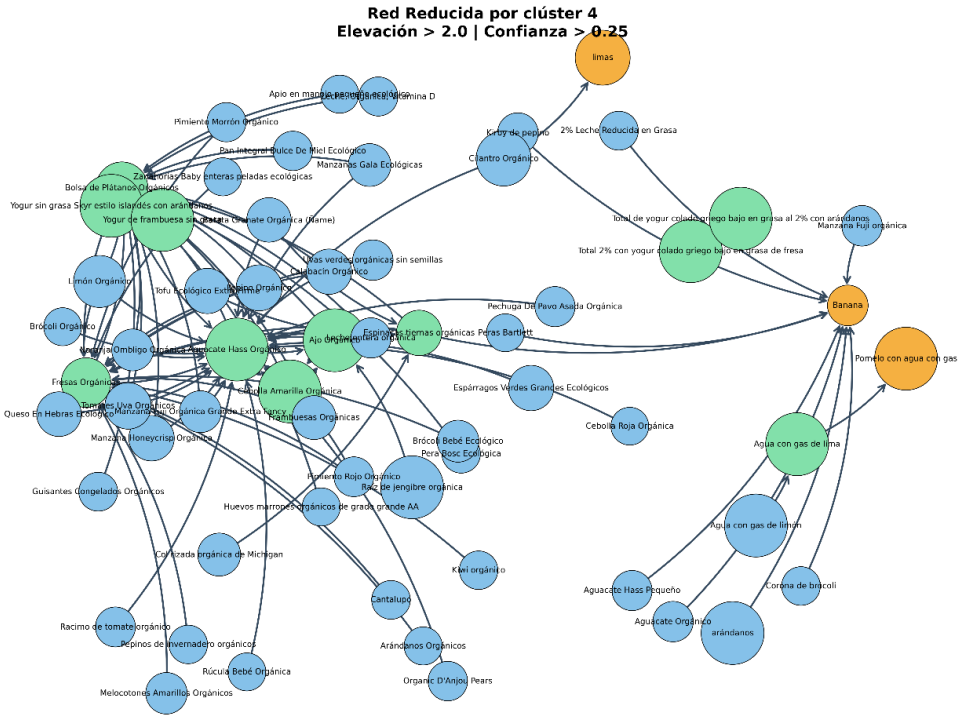

Figura 26: Red de reglas por clúster 4.

En conjunto, las redes de reglas de los clústeres muestran que cada segmento tiene patrones de asociación de productos propios. Aunque la estructura visual varía entre grupos, en todos se observan nodos y enlaces que revelan qué artículos tienden a comprarse juntos dentro de cada perfil. Esto evidencia diferencias claras en las rutas de compra entre clústeres, indicando que cada uno sigue una lógica de consumo particular.

La estrategia de aplicar Apriori segmentado permitió capturar patrones de co-ocurrencia que se diluían al trabajar con la base completa. Las visualizaciones por clúster revelan asociaciones altamente específicas, que pueden alimentar motores de recomendación o decisiones logísticas enfocadas. Esta técnica de análisis cruzado potencia la personalización y evidencia el valor de combinar técnicas de minería de datos descriptiva.

4.5 Visualización final, ejecución local del sistema y enlace a GitHub

Como producto complementario, se desarrolló un dashboard interactivo en Plotly Dash que permite visualizar dinámicamente los principales resultados del análisis. Su estructura incluye menús de navegación lateral, selección de gráficos filtrados por técnica utilizada (Apriori, K-means, cruce de ambas), y visualización integrada de los perfiles de clientes. Este entorno cumple un rol exploratorio, ya que permite revisar los hallazgos del sistema sin necesidad de acceder al código fuente, lo que facilita su comprensión por parte de usuarios no técnicos.

Figura 27: Dashboad k-means, imagen 1.

La ejecución del sistema completo se realiza en entorno local, utilizando un entorno virtual de Python previamente configurado. Una vez descargado el repositorio, basta con activar el entorno, instalar los requerimientos y ejecutar el archivo de lanzamiento (lanzar_dashboard.bat), el cual abre automáticamente la interfaz web. La secuencia de scripts encargados de generar los outputs visuales también se encuentra debidamente organizada por carpetas y numeración temática, permitiendo reproducir el análisis o modificar sus componentes con facilidad.

Este modelo de despliegue asegura la portabilidad del sistema y su posible adaptación a nuevos datasets, manteniendo una arquitectura modular, trazable y documentada.

El repositorio completo del sistema, incluyendo código fuente, estructura modular, datasets procesados, scripts ejecutables y recursos visuales, se encuentra disponible en GitHub:

https://github.com/alfonso-abbott/Proyecto-de-titulo

5- Conclusiones

El proyecto logró desarrollar una solución analítica modular, reproducible y visualmente interpretativa, orientada a descubrir patrones de comportamiento y segmentar clientes de un supermercado online mediante técnicas de minería de datos. Se integraron dos enfoques complementarios: reglas de asociación (Apriori) para identificar productos frecuentemente comprados en conjunto, y clustering (K-means) para agrupar perfiles de clientes. La implementación permitió traducir patrones de datos en conocimiento accionable, evidenciando la aplicabilidad práctica de la analítica descriptiva en contextos comerciales reales.

Conclusión por objetivo específico

1. Explorar el comportamiento de compra mediante análisis de reglas de asociación: Se identificaron asociaciones significativas entre productos mediante Apriori, aplicando métricas de soporte, confianza y lift. Las visualizaciones generadas —como redes, matrices de calor y barras— facilitaron la comprensión de relaciones ocultas en los datos y permitieron detectar agrupaciones de productos relevantes desde el punto de vista comercial.

2. Agrupar clientes en función de sus patrones de compra mediante técnicas de clustering: El uso de K-means permitió segmentar a los clientes en cinco grupos diferenciados, validados con el método del codo y el coeficiente de Silhouette. Las visualizaciones por clúster (distribución, comportamiento de compra, preferencias por departamento y perfil agregado) ofrecieron un mapa interpretativo del ecosistema de clientes.

3. Integrar ambos enfoques para identificar recomendaciones personalizadas por perfil: Mediante el cruce de resultados entre Apriori y K-means, se generaron reglas de asociación específicas por clúster, lo que posibilita futuras estrategias de

recomendación diferenciadas. Esta integración mostró el potencial de combinar técnicas descriptivas para obtener conocimiento accionable más fino y contextualizado.

4. Implementar un entorno modular de desarrollo y visualización local del sistema: Se diseñó un sistema segmentado en scripts, datasets, outputs y dashboard ejecutable localmente, lo que permite una trazabilidad clara del flujo analítico. La solución es portable, ejecutable sin conexión a internet y fácilmente replicable por otros analistas con conocimientos intermedios de Python.

Hallazgos clave

- La mayoría de los productos frecuentemente comprados en conjunto pertenecen a categorías esenciales (despensa, lácteos, frutas).
- Algunos clústeres muestran comportamientos marcadamente nocturnos o diurnos, lo que podría influir en estrategias de marketing por horario.
- El cruce Apriori/K-means permite segmentar no solo al cliente, sino también al contenido de la recomendación.

Impacto técnico

- Se validó la viabilidad de construir sistemas analíticos robustos a partir de datos abiertos y herramientas libres.
- El enfoque modular y documentado favorece su escalabilidad a otros contextos comerciales.
- El uso de visualizaciones avanzadas y dashboards permitió hacer accesibles los resultados a perfiles no técnicos.

Limitaciones

- El dataset base simula compras de un supermercado norteamericano; se requiere validación local para su aplicación directa en Chile.

- No se incluyó un componente predictivo ni de evaluación económica de las recomendaciones.
- Por motivos de tiempo, el dashboard fue desarrollado como producto exploratorio complementario, no como sistema integral de recomendación en tiempo real.

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