

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INDUSTRIAS

**APLICACIÓN DE MODELOS DE MACHINE LEARNING PARA LA TOMA DE
DECISIONES BASADA EN SEGMENTOS DE CLIENTES EN LA INDUSTRIA DE
ALIMENTOS Y BEBIDAS EN CHILE**

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

AUTOR

DAVID VIDAL OSINAGA

PROFESOR GUÍA

PEDRO FERNÁNDEZ DE LA REGUERA B.

PROFESOR CORREFERENTE

GONZALO AMÉSTICA H.

SANTIAGO DE CHILE, AGOSTO, 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Aplicación de modelos de Machine Learning para la toma de decisiones basada en segmentos de clientes en la industria de alimentos y bebidas en Chile

Nombre del candidato(a): David Vidal Osinaga

Carrera / Grado: Ingeniería Civil Industrial

Campus: Santiago, Vitacura Departamento: Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Pedro Fernández De la Reguera Bastidas, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

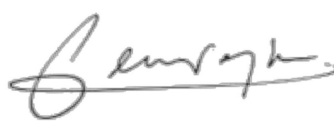
☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 17/08/2026

Firma: 

Estudiante o Candidato(a):

Fecha: 17/08/2026

Firma: 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen Ejecutivo

Esta investigación tiene como propósito analizar cómo las técnicas de Machine Learning pueden apoyar la segmentación de clientes y mejorar la toma de decisiones estratégicas en la industria chilena de alimentos y bebidas .

El estudio surge ante la necesidad de superar los métodos tradicionales de segmentación, basados principalmente en variables demográficas o socioeconómicas, los cuales no siempre permiten comprender la diversidad actual de los consumidores. Para ello, se utilizaron datos de la IX Encuesta de Presupuestos Familiares del INE, integrando una base de personas y una base de cantidades, lo que permitió construir una base final de **13.316 hogares** con información de ingreso, gasto, composición del hogar, productos y canales de compra.

La metodología aplicada fue cuantitativa, descriptiva y aplicada. Primero se realizó la limpieza, integración y transformación de los datos. Luego se construyeron variables relevantes, como gasto total, gasto per cápita, ingreso per cápita, gasto en bebidas, cantidad de productos, número de establecimientos y participación del canal supermercado. Posteriormente, se aplicaron modelos supervisados y no supervisados.

En los modelos supervisados se utilizaron **Regresión Logística, KNN y SVM** para predecir la clasificación socioeconómica de los hogares. Los resultados mostraron un desempeño moderado, destacando SVM con un **Accuracy de 56,32%** y Regresión Logística con el mejor **ROC-AUC macro de 76,23%**. Esto indica que los patrones de consumo entregan señales útiles para diferenciar segmentos, aunque no reemplazan completamente variables directas como ingreso y educación.

En el análisis no supervisado se aplicó **K-Means**, identificando cuatro perfiles de hogares: **familias jóvenes de gasto moderado, hogares mayores de consumo conservador, compradores selectivos de mayor ingreso y consumidores intensivos multicanal**. Estos segmentos permiten comprender mejor los distintos comportamientos de compra y pueden apoyar decisiones de marketing, promociones, surtido, precios y planificación de demanda.

Esta investigación demuestra que el Machine Learning es una herramienta útil para complementar la segmentación tradicional en la industria de alimentos y bebidas. Si bien los modelos predictivos presentan limitaciones, la segmentación obtenida entrega información accionable para la toma de decisiones estratégicas y abre oportunidades para futuras investigaciones con datos transaccionales de empresas, territoriales y de preferencias de consumidores.

Índice de Contenidos

1. Introducción	1
1.1 Problema de Investigación	1
1.2 Objetivos	4
1.2.1 Objetivo general	4
1.2.2 Objetivos específicos	4
1.3 Supuestos de investigación	5
2. Estado del Arte y Alcance	6
2.1 Estado del arte	6
2.1.1 Aplicaciones internacionales	6
2.1.2 Aplicaciones en América Latina	7
2.1.3 Brechas identificadas	8
2.2 Alcance	9
2.2.1 Delimitación geográfica	9
2.2.2 Delimitación temporal	10
2.2.3 Delimitación temática	10
2.2.4 Limitaciones del estudio	11
2.2.5 Alcance general del estudio	12
3. Marco Teórico	12
3.1 Industria de alimentos y bebidas en Chile	12
3.1.1 Relevancia económica del sector	13
3.1.2 Evolución histórica	14
3.1.3 Empleo y cadenas de valor	14
3.1.4 Tendencias de consumo en Chile	15
3.1.5 Competencia y estructura de mercado	15
3.1.6 Retos del sector	16
3.1.7 Oportunidades de innovación con Machine Learning	16
3.2 Segmentación de clientes	17
3.3 Machine Learning en los negocios	18
3.3.1 Definición y evolución	18
3.3.2 Arquitectura del machine learning	19
3.3.3 Training y testing datasets	20

3.3.4 Algoritmos y metodologías de machine learning.....	21
3.3.5 Aprendizaje supervisado.....	21
3.3.6 Aprendizaje No supervisado	26
4. Metodología.....	30
4.1 Enfoque metodológico	30
4.2 Fuentes de datos	31
4.3 Preprocesamiento de los datos	34
4.4 Construcción de los modelos	37
4.4.1 Modelos supervisados	38
4.4.2 Modelos no supervisados	42
4.5 Evaluación de los modelos.....	44
4.5.1 Modelos supervisados	44
4.5.2 Modelos no supervisados	46
4.6 Herramientas utilizadas.....	47
5. Resultados.....	48
5.1 Análisis descriptivo de los datos	48
5.1.1 Matriz de correlación de variables principales	56
5.2 Resultados de modelos supervisados.....	57
5.3 Resultados de modelos no supervisados.....	61
5.4 Análisis de perfiles de clústers	63
5.5 Discusión de hallazgos.....	66
6. Conclusiones.....	67
6.1 Conclusiones generales	67
6.2 Conclusiones específicas.....	68
Referencias.....	69
Anexos	73
Anexo A. Diccionario de variables construidas.....	73
Anexo B. Código Python utilizado explicación por bloques	74
Anexo B.1 Librerías.....	74
Anexo B.2 Configuración de rutas.....	75
Anexo B.3 Funciones auxiliares.....	75
Anexo B.4 Carga de datos.....	77
Anexo B.5 Construcción de variables de personas a nivel hogar.....	78

Anexo B.6 Construcción de variables de consumo a nivel hogar.....	78
Anexo B.7 Integración final de bases y creación de ratios	80
Anexo B.8 Análisis descriptivo de datos	81
Anexo B.9 Gráficos descriptivos del capítulo 5.1	84
Anexo B.10 Modelos Supervisados	85
Anexo B.11 Gráficos de matrices de confusión	88
Anexo B.12 Curva ROC del mejor modelo según ROC-AUC.....	88
Anexo B.13 Modelo No Supervisado K-means	89
Anexo B.14 Evaluación de K-means: Método del codo y Average silhouette	89
Anexo B.15 Entrenamiento final de K-means con K=4.....	90
Anexo B.16 Visualización PCA de clusters	91

Índice de Tablas

Tabla 2.1: Limitaciones del estudio.....	11
Tabla 2.2: Alcance general de la investigación.	12
Tabla 4.1: Resumen de bases de datos utilizadas.....	31
Tabla 4.2. Variables de la base de datos de personas.	32
Tabla 4.3. Variables de la base de datos de cantidades.....	34
Tabla 4.4. Pasos de preprocesamiento realizados.....	36
Tabla 5.1: Estadísticos descriptivos de variables principales de la base integrada.	49
Tabla 5.2: Gasto e ingreso promedio por macrozona, incluyendo cantidad de hogares.....	50
Tabla 5.3: Gasto e ingreso promedio por clasificación socioeconómica, incluyendo cantidad de hogares.....	51
Tabla 5.4: Principales establecimientos por gasto total, registros y cantidad de hogares.	53
Tabla 5.5: Principales categorías por gasto total, registros y cantidad de hogares.....	55
Tabla 5.6: Métricas comparativas de modelos supervisados.....	58
Tabla 5.7: Inercia y Average Silhouette para selección de k.	62
Tabla 5.8: Perfil de clústers obtenidos mediante K-Means.....	65
Tabla A.1: Variables principales construidas.	73

Índice de Figuras

Figura 2.1: Mapa de Chile con cobertura regional de la Encuesta de Presupuestos Familiares (EPF).	9
Figura 2.2: Esquema temático del estudio: variables, relaciones y resultados esperados	10
Figura 3.1: Arquitectura comúnmente usada para el machine learning.....	19
Figura 3.2: Regresión logística de $\pi_i X$ en función de $\beta_0 + \beta_1 x_1, i$	23
Figura 3.3: Clasificación de pacientes según tipo de leucemia mediante el algoritmo de máquinas de vectores de soporte.	26
Figura 3.4: Última iteración del algoritmo de k-medias para un ejemplo de datos sobre	29
Preferencias de consumo de un producto particular.	29
Figura 4.1: Flujo metodológico de la investigación	30
Figura 5.1: Gasto promedio por macrozona.	51
Figura 5.2: Ingreso promedio por macrozona.	52
Figura 5.3: Gasto promedio por clasificación socioeconómica.	52
Figura 5.4: Ingreso promedio por clasificación socioeconómica.	53
Figura 5.5: Gasto total por establecimiento de compra.....	54
Figura 5.6: Categorías por gasto total.	56
Figura 5.7: Matriz de correlación de variables principales.	57
Figura 5.8: Matriz de confusión de Regresión Logística.	59
Figura 5.9: Matriz de confusión - KNN.	60
Figura 5.10: Matriz de confusión - SVM.	60
Figura 5.11: Curva ROC One-vs-Rest del modelo con mayor ROC-AUC macro.	61
Figura 5.12: Método del codo para K-Means.	62
Figura 5.13: Método Average Silhouette para K-Means.	63
Figura 5.14: Visualización PCA de clústers.....	64

1. Introducción

1.1 Problema de Investigación

La industria de alimentos y bebidas en Chile se enfrenta a un entorno cada vez más complejo y competitivo, marcado por cambios en los hábitos de consumo, la digitalización de los canales de venta y la creciente demanda por productos saludables y sostenibles. De acuerdo con el Instituto Nacional de Estadísticas (INE, 2023), este sector representa cerca del 25 % del Producto Interno Bruto (PIB) industrial y genera más de 500.000 empleos directos e indirectos. No obstante, las empresas del sector se ven obligadas a innovar continuamente en sus estrategias comerciales, ya que los métodos tradicionales de segmentación de clientes (basados en edad, género o nivel socioeconómico) resultan insuficientes para capturar la diversidad y dinamismo de los consumidores actuales. La ausencia de enfoques más sofisticados para entender el comportamiento de los clientes puede llevar a decisiones subóptimas en: diseño de productos, estrategias de precios, distribución, logística y campañas de marketing.

Durante las últimas dos décadas, la globalización y la digitalización han transformado radicalmente la forma en que los consumidores adquieren y valoran los alimentos y bebidas. La pandemia de COVID-19 aceleró la adopción de canales digitales, generando un crecimiento del 56 % en las ventas en línea entre 2019 y 2021 (Euromonitor, 2022). Paralelamente, ha emergido un consumidor más informado, preocupado por la salud, la sostenibilidad y la trazabilidad de los productos (FAO, 2022).

Estas transformaciones han impactado directamente en la estructura competitiva del mercado. Grandes conglomerados como Coca-Cola Andina, CCU y Embonor continúan liderando la producción de bebidas, mientras que un número creciente de pequeñas y medianas empresas (PYMES) ingresan al mercado con propuestas de valor diferenciadas en

torno a alimentos orgánicos, veganos o con atributos funcionales (Banco Central de Chile, 2022).

A pesar de la creciente disponibilidad de datos públicos del INE y de la digitalización del sector, no se han documentado estudios sistemáticos que apliquen modelos de Machine Learning para la segmentación de clientes en la industria chilena de alimentos y bebidas. En Chile, la mayoría de los estudios disponibles se concentra en análisis descriptivos del consumo (INE, 2023; MINSAL, 2021), sin integrar modelos predictivos o de clustering. Actualmente, las empresas de este sector enfrentan desafíos en la identificación de patrones de comportamiento de los clientes y la adaptación rápida a las preferencias cambiantes. Se requiere un enfoque más avanzado que aproveche la gran cantidad de datos disponibles para lograr una segmentación precisa y oportuna.

La aplicación de modelos de machine learning ofrece la oportunidad de identificar patrones complejos y dinámicos en los datos del cliente (Ruan y Siau, 2019), lo que puede agilizar el proceso de toma de decisiones (Tecoalu et al., 2021) para impulsar la personalización de productos, el marketing dirigido y la lealtad del cliente. Sin embargo, existe una brecha entre la teoría y la práctica en este campo, lo que justifica una investigación detallada.

Los principales agentes involucrados en esta problemática son las empresas de alimentos y bebidas en Chile que buscan mejorar la segmentación de clientes y los consumidores cuyos comportamientos y preferencias son fundamentales para la segmentación. Esta relación se establece a través de la recopilación de datos de consumo y la implementación de estrategias de marketing y producción basadas en la segmentación.

En este contexto, el problema central de esta investigación consiste en la necesidad de mejorar la toma de decisiones en la industria de alimentos y bebidas en Chile, mediante la aplicación de modelos de machine learning para la segmentación de clientes que, a diferencia de los métodos convencionales, permiten identificar patrones ocultos en grandes volúmenes de datos, predecir comportamientos y descubrir segmentos de clientes más precisos.

En base a lo planteado anteriormente surge la pregunta ¿Cómo pueden los modelos de *Machine Learning* contribuir a la segmentación y predicción de clientes en la industria de alimentos y bebidas en Chile, utilizando datos del INE, para mejorar la toma de decisiones estratégicas? ¿Cuáles son las características más influyentes en la segmentación de clientes en la industria de alimentos y bebidas? ¿Cómo se pueden evaluar y comparar de manera efectiva diferentes modelos de machine learning en la toma de decisiones basadas en la segmentación de clientes?

Se pretende resolver este problema mediante la implementación de modelos de machine learning apoyado en el uso del lenguaje de programación Python, con su entorno de software para realizar el análisis y manipulación de datos, mediante modelos de machine learning de aprendizaje supervisado utilizando los algoritmos de regresión logística, K-Nearest Neighbor y Support Vector Machine, y aprendizaje no supervisado empleando los algoritmos de K-Means, que permita mostrar información relevante para poder implementar mejores estrategias y tomar decisiones con información cuantitativa.

1.2 Objetivos

1.2.1 Objetivo general

Aplicar modelos de Machine Learning para la segmentación y predicción de clientes en la industria de alimentos y bebidas en Chile, utilizando datos del INE, con el fin de optimizar la toma de decisiones estratégicas.

1.2.2 Objetivos específicos

1. Recolectar y procesar bases de datos del INE relacionadas con consumo de alimentos y bebidas.
2. Implementar modelos supervisados (Regresión Logística, KNN y SVM) para predecir patrones de consumo.
3. Aplicar modelos no supervisados (K-Means) para identificar segmentos de clientes.
4. Evaluar y comparar el desempeño de los modelos aplicados.

1.3 Supuestos de investigación

Para desarrollar el análisis y los modelos de *Machine Learning*, se establecen los siguientes supuestos fundamentales:

1. **Representatividad estadística:** Los datos del INE son representativos del comportamiento promedio de los hogares chilenos, conforme a su diseño probabilístico y cobertura nacional (INE, 2023).
2. **Calidad de los datos:** Se asume que los datos recopilados en la EPF (encuesta de presupuestos familiares) presentan un nivel de consistencia y confiabilidad adecuado para análisis estadísticos y modelado predictivo, tras la aplicación de técnicas de limpieza y normalización (Han et al., 2012).
3. **Independencia parcial de las observaciones:** Los registros de los hogares se consideran independientes entre sí, permitiendo aplicar modelos ML basados en aprendizaje supervisado y no supervisado sin dependencia temporal.
4. **Aplicabilidad práctica de los resultados:** Se asume que los hallazgos obtenidos a partir de los modelos pueden trasladarse al ámbito empresarial, contribuyendo a la toma de decisiones estratégicas en marketing y operaciones (Davenport & Bean, 2018).

2. Estado del Arte y Alcance

2.1 Estado del arte

La aplicación de **Machine Learning (ML)** en el ámbito de la segmentación de clientes ha crecido exponencialmente en los últimos veinte años, impulsada por la digitalización de los mercados y la disponibilidad de grandes volúmenes de datos. En particular, la **industria de alimentos y bebidas** ha mostrado un interés creciente en utilizar estas técnicas para **optimizar precios, personalizar estrategias de marketing y anticipar cambios en los patrones de consumo** (Kumar, Dixit, Javalgi, Dass, & Sridhar, 2021).

2.1.1 Aplicaciones internacionales

En países desarrollados, el uso de ML en el consumo de alimentos se ha concentrado en tres áreas principales:

1. Predicción de demanda:

- En Europa, **Gupta y George (2016)** aplicaron modelos de regresión logística y SVM para predecir el consumo de bebidas carbonatadas, logrando una mejora del 18 % en la precisión frente a métodos estadísticos tradicionales.
- En Estados Unidos, se han usado árboles de decisión para pronosticar el impacto de promociones en supermercados, permitiendo ajustar inventarios y evitar sobrestock (Davenport & Bean, 2018).

2. Segmentación de clientes:

- Un estudio realizado en España utilizó **K-Means** para identificar perfiles de consumidores de vino, detectando cinco segmentos diferenciados basados en edad, nivel socioeconómico y frecuencia de consumo (Martínez-López, Gázquez-Abad, & Sousa, 2017).

- En Reino Unido, el uso de clustering jerárquico permitió agrupar consumidores de snacks según patrones de compra en supermercados, lo que facilitó la personalización de ofertas digitales (Hassan, Niazi, & Hussain, 2018).

Estos casos muestran que el ML no solo permite **clasificar y agrupar consumidores**, sino también **mejorar la eficiencia de las decisiones empresariales**.

2.1.2 Aplicaciones en América Latina

En América Latina, aunque el uso de ML en la industria de alimentos y bebidas aún se encuentra en etapas incipientes, existen experiencias relevantes:

- En Brasil, **Silva et al. (2020)** aplicaron modelos de regresión logística y KNN para analizar el consumo de bebidas energéticas entre universitarios, encontrando que variables como la edad y el nivel de ingreso eran altamente predictivas del patrón de consumo.
- En México, **Hernández y Ponce (2019)** usaron clustering jerárquico para segmentar clientes de supermercados con base en la frecuencia de compra y el ticket promedio, identificando cuatro grupos principales que orientaron la estrategia de precios dinámicos.
- En Argentina, un estudio de **Soria y Fernández (2021)** aplicó K-Means a datos de consumo de alimentos saludables, identificando un segmento creciente de consumidores con preferencia por productos orgánicos, lo que reflejó la tendencia global hacia dietas más sostenibles.
- En Chile, aunque los estudios son más limitados, se han realizado aplicaciones de minería de datos en cadenas de supermercados para analizar patrones de consumo regionales (INE, 2023). Estos esfuerzos demuestran el potencial de aplicar ML en la

toma de decisiones estratégicas del sector, particularmente en un país con alta penetración digital y comercio electrónico en alimentos (Euromonitor, 2021).

2.1.3 Brechas identificadas

La revisión de la literatura muestra que:

1. **Internacionalmente**, los algoritmos de ML han demostrado mejoras sustantivas en la predicción de consumo, segmentación de clientes y optimización de estrategias comerciales.
2. **En América Latina**, los estudios son menos frecuentes y suelen concentrarse en sectores de retail general, más que en la industria de alimentos y bebidas específicamente.
3. **En Chile**, existe una clara **brecha de aplicación**. A pesar de la disponibilidad de datos a través del INE y la digitalización de canales de compra, no se han documentado investigaciones sistemáticas que integren ML para la segmentación de clientes en este sector.

En consecuencia, la presente investigación busca **llenar ese vacío**, contribuyendo con evidencia empírica sobre la utilidad de los modelos supervisados y no supervisados en la industria chilena de alimentos y bebidas.

2.2 Alcance

2.2.1 Delimitación geográfica

La investigación se circunscribe al **territorio nacional de Chile**, abarcando los hogares participantes de la *Encuesta de Presupuestos Familiares (EPF)* elaborada por el **Instituto Nacional de Estadísticas (INE)**.

Esta encuesta posee representatividad en las **16 regiones** del país, con una cobertura que incluye áreas urbanas y rurales, lo que la convierte en una fuente válida para analizar el consumo de alimentos y bebidas de los hogares chilenos (INE, 2021-2022).

El estudio se enfoca principalmente en las **zonas urbanas**, donde la diversidad de canales de compra (supermercados, minimarkets, plataformas online y ferias libres) permite observar la interacción entre el comportamiento tradicional y el digital.

Figura 2.1: *Mapa de Chile con cobertura regional de la Encuesta de Presupuestos Familiares (EPF). Fuente: Elaboración propia con base en la IX EPF INE (2021-2022).*

Cobertura territorial de la muestra integrada



El estudio no busca analizar diferencias comunales o microterritoriales, sino establecer patrones generales de consumo a nivel regional y nacional.

2.2.2 Delimitación temporal

El período de análisis se centra en dos horizontes de tiempo complementarios:

1. **Base primaria:** la *IX Encuesta de Presupuestos Familiares (EPF Octubre 2021–Septiembre 2022)*, que proporciona datos individuales de gasto, ingreso y hábitos de consumo por hogar.

El análisis temporal no pretende realizar comparaciones longitudinales, sino **proporcionar una fotografía representativa y actualizada del consumo de alimentos y bebidas en Chile**.

En consecuencia, el estudio combina información histórica estructurada (EPF) para contextualizar los resultados de los modelos de *Machine Learning*.

2.2.3 Delimitación temática

El alcance temático se centra en el **consumo de alimentos y bebidas** en los hogares chilenos, considerando variables cuantitativas que permitan comprender los patrones de gasto y comportamiento del consumidor.

Figura 2.2: Esquema temático del estudio: variables, relaciones y resultados esperados

Fuente: Elaboración propia.



El estudio no incluye variables de tipo **psicográfico** (valores, estilo de vida, motivaciones), debido a la ausencia de datos en la EPF, ni tampoco variables transaccionales privadas de empresas del retail o plataformas digitales.

2.2.4 Limitaciones del estudio

A pesar del rigor metodológico, el estudio presenta limitaciones que deben reconocerse para interpretar adecuadamente los resultados.

Tabla 2.1: Limitaciones del estudio. Fuente: Elaboración propia.

Tipo de limitación	Descripción	Posible impacto
Temporal	La EPF se actualiza aproximadamente cada 4 años, lo que puede no reflejar cambios recientes en hábitos de consumo post-pandemia.	Menor vigencia de algunos patrones de gasto.
Temática	No incluye variables de preferencias personales, psicográficas ni actitudinales.	Menor profundidad en el perfilamiento de consumidores.
Geográfica	No se realiza análisis comunal ni microsegmentación por ciudades.	Limitación en la precisión local de los segmentos.
Técnica	Los modelos ML se aplican sobre datos agregados por hogar, no por individuo.	Reducción de granularidad y precisión predictiva.

Estas limitaciones no invalidan la investigación, pero sí definen un marco realista sobre los alcances de generalización de los resultados.

2.2.5 Alcance general del estudio

En síntesis, el alcance de la investigación se define según los siguientes ejes:

Tabla 2.2: Alcance general de la investigación. Fuente: Elaboración propia.

Dimensión	Alcance definido	Fuentes de referencia
Geográfica	Chile, con cobertura nacional (zonas urbanas y rurales).	INE (2023)
Temporal	EPF 2021-2022.	INE, FAO
Temática	Consumo y gasto en alimentos y bebidas.	INE, MINSAL
Analítica	Aplicación de ML supervisado y no supervisado.	Hastie et al. (2009)
Instrumental	Uso de Python, Scikit-learn, Pandas y Matplotlib.	Han et al. (2012)

3. Marco Teórico

3.1 Industria de alimentos y bebidas en Chile

La **industria de alimentos y bebidas en Chile** constituye uno de los pilares de la economía nacional, tanto por su relevancia en el consumo interno como por su papel en las exportaciones. De acuerdo con el **Instituto Nacional de Estadísticas (INE, 2023)**, este sector representa aproximadamente un 25 % del valor agregado industrial, generando un importante volumen de empleo y encadenamientos productivos.

En el ámbito local, el consumo de alimentos y bebidas ha evolucionado en las últimas décadas hacia patrones cada vez más diversificados, influenciados por factores como la globalización, la urbanización, los cambios en los ingresos de los hogares y las

transformaciones culturales (FAO, 2022). Así, se observa un crecimiento en la demanda de productos procesados y de conveniencia, así como un aumento en el interés por opciones más saludables y sostenibles (Euromonitor, 2021).

Por otra parte, el sector enfrenta desafíos estructurales. La **concentración de la oferta** en grandes cadenas de supermercados y distribuidoras limita las oportunidades de pequeños productores, mientras que la creciente digitalización de los canales de venta introduce nuevos mecanismos de competencia (INE, 2023). Este escenario obliga a las empresas a innovar en sus estrategias de marketing y distribución para captar la atención de consumidores cada vez más informados y exigentes.

En este contexto, la **segmentación de clientes** basada en datos emerge como una herramienta estratégica para identificar patrones de consumo, anticipar tendencias y diseñar productos o servicios ajustados a las necesidades de cada segmento. Sin embargo, las técnicas tradicionales de segmentación resultan insuficientes para abordar la complejidad actual, lo que abre espacio al uso de **herramientas analíticas avanzadas como el Machine Learning**.

3.1.1 Relevancia económica del sector

La **industria de alimentos y bebidas en Chile** es uno de los sectores más relevantes dentro de la manufactura nacional. Según el **Instituto Nacional de Estadísticas (INE, 2023)**, representa cerca de un **25 % del Producto Interno Bruto (PIB) industrial** y aporta aproximadamente un **12 % del total del empleo manufacturero**.

De acuerdo con datos del **Banco Central de Chile (2022)**, las exportaciones de alimentos superaron los **US\$ 20.000 millones** anuales, posicionando al país como uno de los principales proveedores de frutas, vinos, salmones y bebidas en mercados internacionales. En particular, el vino chileno ha alcanzado una participación destacada en la Unión Europea

y China, mientras que el salmón constituye uno de los productos estrella hacia Estados Unidos y Japón.

Este dinamismo convierte al sector en un **motor estratégico de la economía nacional**, no solo por sus ingresos, sino también por su capacidad de generar encadenamientos productivos con la agricultura, la pesca y la logística.

3.1.2 Evolución histórica

El crecimiento de la industria de alimentos en Chile puede analizarse en tres etapas:

1. **Décadas de 1970–1990:** marcada por la apertura comercial y la consolidación de exportaciones de productos primarios como frutas, vinos y mariscos (Meller, 1996).
2. **Década de 2000:** diversificación hacia productos con mayor valor agregado, como jugos concentrados, procesados y bebidas funcionales (Agosin, 2007).
3. **2010 en adelante:** transición hacia alimentos saludables, orgánicos y sustentables, en línea con tendencias internacionales de consumo responsable (FAO, 2022).

La industria ha transitado desde un modelo centrado en **productos básicos para exportación** hacia un esquema donde la **innovación, el marketing y la diferenciación** son claves para mantener la competitividad.

3.1.3 Empleo y cadenas de valor

El sector emplea aproximadamente a **500.000 trabajadores directos e indirectos**, lo que lo convierte en uno de los mayores generadores de empleo dentro de la manufactura chilena (INE, 2023). Además, promueve la participación de pequeñas y medianas empresas (PYMES), las cuales tienen un rol relevante en la producción y distribución local de alimentos y bebidas.

La cadena de valor del sector se articula en distintos eslabones:

- **Producción primaria:** agricultura, pesca y ganadería.
- **Procesamiento:** industrias de bebidas, lácteos, carnes y productos envasados.
- **Distribución:** supermercados, minimarkets, plataformas digitales.
- **Consumo final:** hogares y canales HORECA (hoteles, restaurantes y catering).

Cada eslabón genera información valiosa que puede ser analizada mediante **herramientas de Machine Learning** para mejorar la eficiencia y personalizar la oferta.

3.1.4 Tendencias de consumo en Chile

Los cambios en los hábitos de consumo de los chilenos reflejan tendencias globales:

- **Mayor interés en alimentos saludables:** El **Ministerio de Salud (MINSAL, 2021)** reporta un aumento sostenido en la demanda de productos bajos en sodio, azúcar y grasas.
- **Crecimiento del comercio electrónico:** Según **Euromonitor (2021)**, las ventas online de alimentos y bebidas crecieron un **56 % entre 2019 y 2021**, aceleradas por la pandemia de COVID-19.
- **Sustentabilidad y trazabilidad:** Estudios muestran que un segmento creciente de consumidores chilenos está dispuesto a pagar más por productos con certificaciones ambientales y sociales (FAO, 2022).
- **Preferencia por conveniencia:** Aumento en el consumo de comidas listas para servir y bebidas funcionales, especialmente entre jóvenes urbanos.

Estas tendencias complejizan la **segmentación de clientes**, ya que las preferencias se diversifican y fragmentan más allá de criterios demográficos tradicionales.

3.1.5 Competencia y estructura de mercado

El mercado chileno de alimentos y bebidas está dominado por **grandes cadenas de supermercados** como Cencosud, Walmart Chile y SMU, que concentran más del 60 % de

las ventas (Euromonitor, 2021). Sin embargo, también se observa una expansión de **emprendimientos locales y marcas nicho**, que compiten con propuestas diferenciadas de alimentos orgánicos, gourmet o veganos.

En cuanto a bebidas, destacan multinacionales como **Coca-Cola Andina, CCU y Embonor**, que concentran la mayoría del mercado de bebidas carbonatadas y cervezas. Estas empresas han invertido en **analítica avanzada y segmentación de clientes** para enfrentar la creciente competencia de marcas artesanales y bebidas saludables.

3.1.6 Retos del sector

A pesar de su relevancia, el sector enfrenta varios desafíos:

- **Volatilidad de precios internacionales:** La dependencia de mercados externos expone al sector a fluctuaciones en la demanda global.
- **Normativas de etiquetado y salud pública:** La Ley de Etiquetado de Alimentos en Chile (2016) obligó a reformular productos, lo que implicó altos costos para la industria.
- **Cambio en preferencias del consumidor:** El tránsito hacia productos más saludables y sustentables obliga a una **adaptación constante de la oferta**.
- **Competencia digital:** La entrada de plataformas internacionales de e-commerce desafía a los actores locales a innovar en logística y distribución.

3.1.7 Oportunidades de innovación con Machine Learning

El sector de alimentos y bebidas en Chile tiene un **alto potencial para aprovechar el Machine Learning** en múltiples áreas:

- **Segmentación avanzada de clientes:** Identificar perfiles de consumo ocultos en datos de gasto y hábitos de compra.
- **Predicción de demanda:** Anticipar necesidades de inventario para reducir mermas.
- **Optimización de precios:** Ajustar precios dinámicamente según características del cliente y contexto.
- **Recomendaciones personalizadas:** Usar algoritmos de ML para sugerir productos a consumidores en plataformas digitales.

El aprovechamiento de estas herramientas permitirá a las empresas locales **competir de manera más efectiva en un entorno globalizado** y responder con mayor agilidad a las cambiantes preferencias de los consumidores.

3.2 Segmentación de clientes

La segmentación de clientes se define como el proceso de dividir un mercado heterogéneo en grupos más homogéneos de consumidores que comparten características similares, con el fin de diseñar estrategias diferenciadas de marketing (Kotler & Keller, 2016). Tradicionalmente, la segmentación se ha realizado con base en variables demográficas (edad, género), geográficas (región, ciudad), psicográficas (valores, estilos de vida) y conductuales (frecuencia de compra, fidelidad a la marca).

No obstante, estos enfoques presentan **limitaciones**. En primer lugar, suponen que las características observables de los consumidores son suficientes para predecir su comportamiento, lo cual rara vez se cumple en mercados dinámicos (Wedel & Kamakura, 2012). En segundo lugar, no capturan la interacción entre múltiples variables ni permiten detectar patrones ocultos en grandes volúmenes de datos.

En contraste, la **segmentación basada en datos** utiliza técnicas estadísticas y de inteligencia artificial para identificar grupos de consumidores a partir de su comportamiento real. Por ejemplo, en la industria de alimentos y bebidas, es posible agrupar clientes en función de sus patrones de gasto en productos saludables, bebidas alcohólicas o alimentos procesados. Esto permite a las empresas diseñar campañas de marketing más personalizadas y eficaces (Wedel & Kamakura, 2012; Kumar et al., 2021).

En la literatura reciente se ha documentado que la segmentación basada en Machine Learning supera en precisión a los métodos tradicionales, al permitir que los algoritmos aprendan automáticamente estructuras ocultas en los datos (Ascarza, 2018). Este enfoque se alinea con la tendencia global hacia el **marketing basado en datos (data-driven marketing)**, donde las decisiones estratégicas se sustentan en modelos predictivos.

3.3 Machine Learning en los negocios

3.3.1 Definición y evolución

El **Machine Learning (ML)** se entiende como una rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y mejorar su rendimiento en tareas específicas sin necesidad de ser programados explícitamente (Bishop, 2006).

Su objetivo es construir modelos capaces de **predecir, clasificar o agrupar observaciones** a partir de características observadas.

Matemáticamente, un modelo de ML busca aproximar una función:

$$f: X \rightarrow Y$$

donde X representa las variables predictoras (inputs) e Y la variable objetivo (output).

Su aplicación en los negocios se ha consolidado en las últimas dos décadas gracias al incremento en la disponibilidad de datos masivos y a la capacidad computacional.

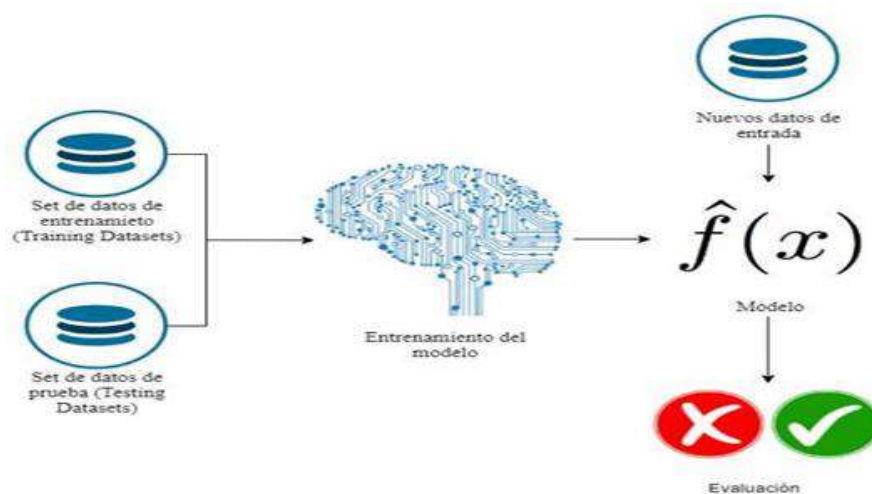
3.3.2 Arquitectura del machine learning

“El flujo de trabajo común para la implementación de un algoritmo de aprendizaje de máquina para la solución de un problema se puede simplificar en 3 etapas principalmente” (Contreras, 2019):

- Primera etapa: Con el set de datos inicial, hago una partición arbitraria de los datos en un set de datos de entrenamiento (training datasets) y en un set de datos de prueba (testing datasets).
- Segunda etapa: Con estos sets de datos establecidos hago el entrenamiento del modelo para que aprenda sobre el comportamiento de los datos.
- Tercera etapa: Con el modelo entrenado se obtiene el modelo final, que buscará predecir el comportamiento de algún nuevo dato de entrada que se le quiera consultar al modelo, para posteriormente entregar información que será evaluada para indicarle así al modelo si se está equivocando o no.

Las etapas mencionadas anteriormente se pueden simplificar en el siguiente diagrama:

Figura 3.1: *Arquitectura comúnmente usada para el machine learning. Fuente: (Contreras, 2019)*



3.3.3 Training y testing datasets

Los datos con los cuales se aplicará el algoritmo de aprendizaje de máquina se le conoce como Set de Datos de Entrenamiento o “Training Datasets”, que consisten en un set de pares (x,y) , Contreras señala, “que este par de datos representan valores y etiquetas” (2019):

- x : es un vector de valores, generalmente llamado vector de características. Cada valor o característica puede ser categórica (es decir, variables cualitativas que representen categorías como por ejemplo colores (negro, blanco, azul), género (masculino y femenino) entre otras categorías) o numéricas (valores enteros o números reales).
- y : es la etiqueta o “label” del valor de clasificación para x .

El objetivo del proceso de aprendizaje de máquina es descubrir la función $y = f(x)$ que pueda predecir de la mejor manera posible el valor de “ y ” asociado con cada valor del vector “ x ”. Dependiendo del valor de “ y ”, que en principio puede ser arbitrario, se pueden encontrar varios casos a considerar (Contreras, 2019):

1. Si “ y ” es un número real, el problema de aprendizaje de máquina se le conoce como regresión.
2. Si “ y ” es un valor dicotómico, como por ejemplo un valor verdadero/falso, si/no, 0/1, el problema de aprendizaje de máquina se le conoce como clasificación binaria.

Ambos casos considerados corresponden a un tipo de aprendizaje supervisado.

Por otro lado, el Set de Datos de Prueba o “Testing Datasets” corresponden a una retención de un porcentaje de los datos de entrenamiento, esto con la finalidad de poder calcular el desempeño del modelo y verificar si el clasificador posee problemas de sobreajuste, y realizar las correcciones necesarias en la construcción del modelo.

3.3.4 Algoritmos y metodologías de machine learning

González señala que, “los algoritmos de Aprendizaje de Máquina se pueden dividir en 2 grupos principales a partir de la utilidad que se requiera (2019):

- **Aprendizaje supervisado**, en el cual los algoritmos se entrenan con datos etiquetados para predecir variables de interés.
- **Aprendizaje no supervisado**, donde los algoritmos buscan estructuras subyacentes en datos no etiquetados.

Ambos enfoques resultan relevantes para la industria de alimentos y bebidas, ya que permiten tanto **predecir patrones de compra** como **descubrir segmentos de consumidores**.

3.3.5 Aprendizaje supervisado

El aprendizaje supervisado comienza típicamente con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos. El aprendizaje supervisado tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de analítica. Estos datos tienen características etiquetadas que definen el significado de los datos (González, 2019).

En los algoritmos de aprendizaje supervisado se genera un modelo predictivo, basado en datos de entrada y salida. La palabra clave “supervisado” viene de la idea de tener un conjunto de datos previamente etiquetado y clasificado, es decir, tener un conjunto de muestra el cual ya se sabe a qué grupo, valor o categoría pertenecen los ejemplos (González, 2019).

Con este grupo de datos, el cual llamamos datos de entrenamiento, se realiza el ajuste al modelo inicial planteado. De esta forma es como el algoritmo va “aprendiendo” a clasificar las muestras de entrada comparando el resultado del modelo, y la etiqueta real de la muestra,

realizando las compensaciones respectivas al modelo de acuerdo a cada error en la estimación del resultado (González, 2019).

Tal como indica González, “los métodos y algoritmos que se pueden implementar para el tipo de aprendizaje supervisado son los siguientes” (2019):

- Regresión logística.
- K-Nearest Neighbor.
- Support Vector Machine.

3.3.5.1 Regresión Logística

La regresión logística es un algoritmo más adecuado para predecir la clase de una variable de carácter categórico en función de los valores que tomen otras variables que dependen de ella. La estimación del valor de la variable dependiente es realizada al escoger la clase más probable en función de los valores que tomen otras variables independientes (Hosmer Jr et al., 2013):

$$Y_i = X\beta = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} \quad (1)$$

Luego se define una función probabilística tal que:

$$P(Y = 1|X) = p_i(X) = \frac{e^{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i}}}{1 + e^{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i}}} \quad (2)$$

donde:

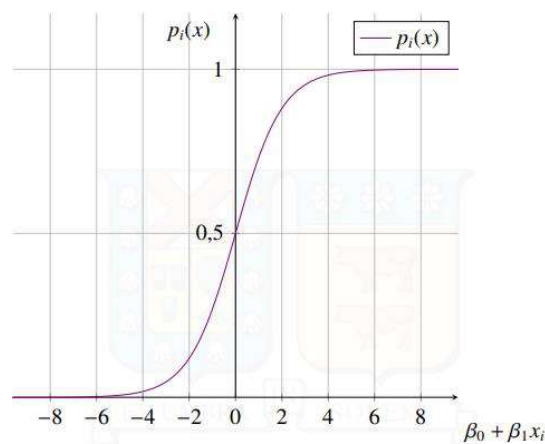
- $P(Y = 1 | X)$ es la probabilidad de que ocurra el evento (por ejemplo, realizar compra online),
- β_0 es el intercepto,
- β_i son los coeficientes estimados para cada variable X_i .

Llevando las ecuaciones 1 y 2 a una regresión logística de una dimensión, es decir, calculando la probabilidad de una variable categórica dado el valor que tome una variable independiente x (Ver ecuaciones 3 y 4), es posible graficar la regresión logística $p_i(X)$ en función de Y_i (Ver Figura 3.2):

$$Y_i = X\beta = \beta_0 + \beta_1 x_{1,i} \quad (3)$$

$$P(Y = 1|X) = p_i(X) = \frac{e^{\beta_0 + \beta_1 x_{1,i}}}{1 + e^{\beta_0 + \beta_1 x_{1,i}}} \quad (4)$$

Figura 3.2: Regresión logística de $p_i(X)$ en función de $\beta_0 + \beta_1 x_{1,i}$. Fuente: (Caba, 2020)



3.3.5.2 K-Nearest Neighbor

El algoritmo de K-Nearest Neighbors (KNN) tiene como principal objetivo identificar un número adecuado de observaciones u objetos que formen parte del conjunto de datos de entrenamiento del modelo que se encuentren más cercanos en términos de distancia a la nueva observación u objeto que se busque estimar o predecir, y en función de los atributos de sus vecinos más cercanos estimar cual es la clase a la que se le debería asignar a la nueva observación. Por lo general, este algoritmo se utiliza para clasificar datos que se caractericen por tener atributos con valores continuos (Bramer, 2007), por lo que es importante mencionar

que se debe estimar correctamente que tipo de medida de distancia es más adecuada utilizar según las características de los datos y del problema.

Es preciso escoger un valor adecuado para el parámetro K tal de obtener resultados que generen valor. Por lo general, un valor para el parámetro K muy pequeño implica que los resultados podrían no entregar valor como se espera si el modelo arroja resultados con mucha variación entre los datos y con presencia de datos atípicos. Por otra parte, para un parámetro K demasiado alto podría disminuir el número de errores del modelo pero a la vez suavizar demasiado la forma en que el modelo estima la clasificación de las observaciones, pues el número de observaciones a considerar para la predicción aumenta (Aggarwal, 2015).

El algoritmo de K-Nearest Neighbors (KNN) no especifica de forma explícita una función que permita clasificar nuevos datos a partir de los datos de entrenamiento del modelo, si no más bien actúa sobre el conjunto de datos de entrenamiento completo al almacenarlo o memorizarlo, y según los atributos de sus dichos datos el modelo genera como respuesta una predicción respecto a la clase de un objeto u observación particular (Runkler, 2012).

3.3.5.3 Support Vector Machine

Las máquinas de vectores de soporte consisten en una serie de algoritmos propios del aprendizaje supervisado, destinado a la clasificación y regresión de datos.

La idea principal llevada a cabo por este algoritmo consiste en lo siguiente (Cortes y Vapnik, 1995):

1. Se comienza el análisis de los datos en sus dimensiones originales, donde cada observación pertenece a una de dos posibles categorías. Dado esto, se busca construir un modelo que sea capaz de clasificar a nuevas observaciones.

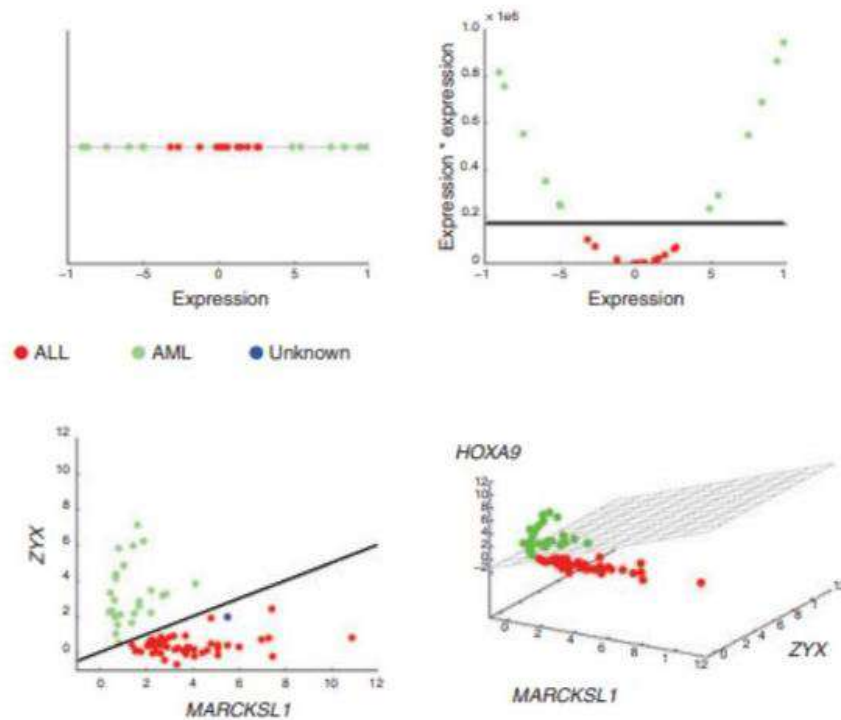
2. El algoritmo estudia las relaciones existentes entre los datos en espacios dimensionales superiores.
3. El algoritmo busca el espacio dimensional más óptimo para calcular las relaciones existentes entre cada par de observaciones, y hace uso de dichas relaciones para escoger el clasificador de vectores de soporte que mejor separe a los datos según sus categorías o clases.

De esta manera, cuando se posee una variable que se desea clasificar según dos categorías, pero no se posee un clasificador lineal que sea capaz de categorizar sus muestras de manera correcta, el algoritmo de máquinas de vectores de soporte es capaz de llevar los datos a un espacio matemático de una dimensión superior que pueda clasificar las observaciones efectivamente.

Para realizar el cálculo de las relaciones existentes entre cada par de observaciones se hace uso de funciones kernel, las cuales se encargan de hallar clasificadores de vectores de soporte sistemáticamente. Las funciones kernel calculan relaciones entre cada par de observaciones al proyectar los datos en un espacio dimensional más alto, pero sin la necesidad de llevar a cabo una transformación matemática de los datos a otros espacios dimensionales (Schölkopf, 2001).

Con el objetivo de ejemplificar el trabajo llevado a cabo por las funciones kernel para proyectar los datos en otros espacios dimensionales, se cita el estudio llevado a cabo por Noble (2006) , donde se clasifica mediante máquinas de vectores de soporte a pacientes con leucemia linfoblástica (ALL) y pacientes con leucemia mieloide (AML):

Figura 3.3: Clasificación de pacientes según tipo de leucemia mediante el algoritmo de máquinas de vectores de soporte. En los gráficos superiores, los datos son separados llevándolos a una dimensión superior al elevar su expresión al cuadrado. En los gráficos inferiores, las observaciones se encuentran separadas a excepción de una observación de clase desconocida, la cual es revelada al llevar los datos a una dimensión. Fuente: (Noble, 2006)



3.3.6 Aprendizaje No supervisado

El aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar. Por ejemplo, las aplicaciones de redes sociales, tales como Twitter, Instagram y Snapchat, tienen grandes cantidades de datos sin etiquetar. La comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con base en los patrones o clústeres que encuentra (González, 2019).

Los algoritmos de aprendizaje no supervisado trabajan de forma muy similar a los supervisados, con la diferencia de que éstos solo ajustan su modelo predictivo tomando en cuenta los datos de entrada, sin importar los de salida. Es decir, a diferencia del supervisado, los datos de entrada no están clasificados ni etiquetados, y no son necesarias estas características para entrenar el modelo. Dentro de este tipo de algoritmos, el agrupamiento o clustering en inglés, es el más utilizado, ya que particiona los datos en grupos que posean características similares entre sí (González, 2019).

Tal como indica González, “los métodos y algoritmos que se pueden implementar para el tipo de aprendizaje No supervisado son los siguientes” (2019):

- K-Medias.
- Agrupamiento jerárquico.

Ambos enfoques son especialmente relevantes para la segmentación de clientes, ya que permiten descubrir perfiles de consumidores no evidentes a partir de variables demográficas tradicionales.

3.3.6.1 K-Medias

El clustering de k-medias es un algoritmo capaz de trabajar con grandes cantidades de datos que busca separar los datos en k particiones, donde cada observación de los datos forma parte del clúster que posea la media más cercana (MacQueen et al., 1967). Cabe destacar que a medida que el algoritmo realiza iteraciones para buscar la solución óptima, todas las observaciones del total de los datos son divididas tal que cada observación sea asignada al menos a uno de los clústers (Kudyba, 2014).

Para un conjunto cualquiera de centroides $z \in Z$, cada uno cuenta con una vecindad definida como el subconjunto de observaciones que se encuentran más cercanas de un

centroide z dado que de cualquier otro centroide perteneciente a Z (Laszlo et al., 2007). Tomando esto en cuenta, el algoritmo se lleva a cabo realizando la siguiente iteración (MacQueen et al., 1967):

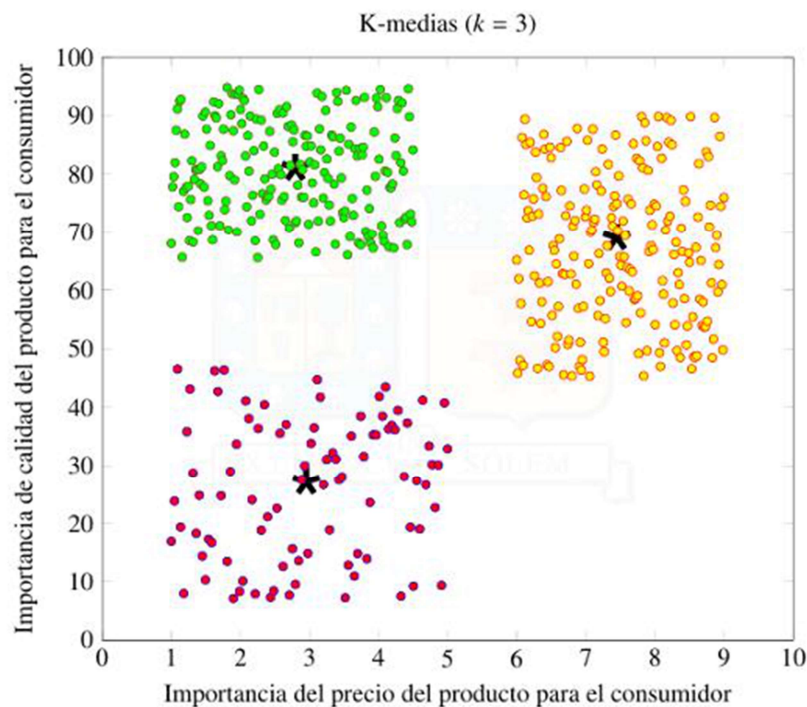
1. Se seleccionan k observaciones iniciales para ser los centroides de los primeros cluster de datos con que comience el algoritmo.
2. Cada observación de los datos es atribuida al clúster de datos que posea el centroide más cercano a cada observación.
3. Ya con los clústers definidos y sus respectivas observaciones asignadas, se procede a calcular nuevamente los valores de los centroides k .
4. Se repiten los pasos 2 y 3 hasta que los k centroides que el algoritmo elija no cambien tras una nueva iteración. Este paso permite minimizar la variabilidad de las observaciones que se encuentren en un mismo clúster y maximizar la distancia entre clústers (Kudyba, 2014).

Para un uso eficaz de este algoritmo en el proceso de extracción de conocimiento, Kudyba (2014) destaca las siguientes consideraciones:

- Se debe hacer una limpieza exhaustiva de los datos atípicos que podrían hallarse en el conjunto de datos a analizar, puesto que tienden a ser dominantes ante el resto de las observaciones en la formación de cada clúster.
- En cuanto a la selección de variables que formen parte del modelo, se debe procurar que estas sean representativas del escenario de negocios y del problema que se plantee, y que dichas variables sean independientes entre sí.
- Respecto a las mediciones que se realicen en torno al algoritmo, es de gran importancia que las variables seleccionadas posean una escala de medida que

permita comparar sus valores y dichos valores deben estar estandarizados tal que puedan ser comparados de manera correcta.

Figura 3.4: Última iteración del algoritmo de k -medias para un ejemplo de datos sobre Preferencias de consumo de un producto particular. Fuente: (Caba, 2020)



Se debe destacar que el uso de este algoritmo no garantiza que converja hacia el óptimo global al tratarse de algoritmo heurístico, por lo que los resultados que este entregue son muy sensibles a los centroides iniciales que sean escogidos, donde si estos centroides iniciales son mal escogidos el algoritmo podría derivar en una solución muy alejada del óptimo global (Laszlo y Mukherjee, 2007). Dado que el algoritmo puede ejecutarse de forma indefinida y de manera rápida, es posible escoger entre varias opciones de centroides iniciales hasta que uno de dichos conjuntos permita entregar una solución óptima y que aporte valor para la toma de decisiones.

4. Metodología

4.1 Enfoque metodológico

La presente investigación sigue un **enfoque cuantitativo, aplicado y exploratorio descriptivo**, dado que se busca analizar datos reales de consumo de alimentos y bebidas en Chile, provenientes del **Instituto Nacional de Estadísticas (INE)**, mediante la aplicación de técnicas de **Machine Learning supervisado y no supervisado**.

Es cuantitativo porque utiliza variables numéricas de ingreso, gasto, cantidad y composición del hogar; aplicado porque busca transformar los resultados en recomendaciones para la industria; y exploratorio-descriptivo porque identifica patrones de consumo y segmentos sin asumir previamente su estructura.

El nivel de análisis es el hogar. Esta decisión metodológica es consistente con la estructura de la EPF, donde la información de gasto se registra por vivienda/hogar y la información sociodemográfica se obtiene desde la base de personas. La unidad de modelamiento final corresponde a hogares integrados que poseen registros válidos en ambas bases.

El flujo metodológico puede resumirse en cinco fases principales (Figura 1):

Figura 4.1: *Flujo metodológico de la investigación. Fuente: Elaboración propia.*



4.2 Fuentes de datos

Las fuentes de datos utilizadas fueron dos archivos CSV de la IX Encuesta de Presupuestos Familiares (EPF): `base-personas-ix-epf.csv` y `base-cantidades-ix-epf.csv`. Ambas bases fueron cargadas con separador punto y coma y separador de miles representado por punto, formato presente en variables como gasto, cantidad e ingreso.

La base de personas contiene variables de caracterización demográfica y económica: macrozona, folio_v, clasificación socioeconómica (cse), número de personas del hogar, parentesco, sexo, edad, estado civil, sustentador principal, escolaridad, tipo de institución laboral, clasificación ocupacional e ingreso disponible. La base de cantidades contiene registros de bienes adquiridos, incluyendo cantidad, unidad de medida, glosa CCIF, gasto y establecimiento de compra.

La integración por folio_v permitió pasar desde registros individuales y registros de productos hacia una base única de hogares. Esta base integrada fue la usada tanto para el **análisis descriptivo** como para los modelos de **Machine Learning**.

Tabla 4.1: Resumen de bases de datos utilizadas. Fuente: Elaboración propia.

Base	Registros	Hogares únicos	Uso Metodológico
Base de personas	44.486	14.851	Caracterización sociodemográfica y económica.
Base de cantidades	122.723	13.316	Gasto, cantidad, productos y canales de compra
Base integrada final	13.316	13.316	Modelamiento supervisado y no supervisado

Variables Principales

Tabla 4.2. Variables de la base de datos de personas. Fuente: Elaboración propia.

Variable	Etiqueta	Valores	Etiqueta valores	Tipo de variable	Formato
macrozona	Macrozona	1 2 3 4	Norte Gran Santiago Centro Sur	Discreta	Numérica
folio_v	Identificador de vivienda	Rango: 1-14961	Válidos	Discreta	Numérica
cse	Clasificación socioeconómica	1 2 3	Bajo Medio Alto	Discreta	Numérica
npersonas	Número de personas en el hogar	Rango: 1-16	Válidos	Discreta	Numérica
parentesco	Parentesco con la persona sustentadora de hogar	1	Sustentador/a del hogar	Discreta	Numérica
		2	Espos/a o conviviente		
		3	Hija/o de ambos		
		4	Hija/o solo de sustentador/a		
		5	Hija/o solo de esposo/a o conviviente		
		6	Hermano/a		
		7	Padre/Madre		
		8	Cuñado/a		
		9	Suegro/a		
		10	Nuera/yerno		
		11	Abuelo/a		
		12	Nieto/a		
		13	Otra persona pariente		
		14	Otra persona no pariente		
sexo	Sexo de la persona	1	Hombre	Discreta	Numérica
		2	Mujer		
		-88 -99	No responde No sabe		
edad	Edad de la persona	Rango: 0-105	Válidos	Discreta	Numérica

ecivil	Estado civil de la persona	1	Casado/a o conviviente	Discreta	Numérica
		2	por Acuerdo de Unión Civil		
		-77	Otro		
		-88	No aplica		
		-99	responder		
sprincipal	Persona sustentadora principal	0	No responde	Discreta	Numérica
		1	No sabe		
			No es el sustentador/a principal		
edue	Años de escolaridad de la persona	Rango: 0 - 24	Sustentador/a principal	Discreta	Numérica
ae12	Tipo de negocio, empresa o institución donde trabaja	1	Válidos	Discreta	Numérica
		2	Estatal		
		3	Privada		
		-77	Hogar particular		
		-88	No aplica		
t_tipo	Clasificación de la ocupación principal	-99	No responde	Discreta	Numérica
		1	No sabe		
		2	Asalariado		
		3	Honorario		
		4	Profesional independiente		
ing_disp_hd	Ingreso disponible imputado por HD	-77	Cuenta Propia	Continua	Numérica
			No aplica		
			responder		
ing_disp_hog_hd	Ingreso disponible del hogar imputado por HD	Rango: -1167072 - 24830041	Válidos	Continua	Numérica
		Rango: -500000 - 34590449	Válidos	Continua	Numérica

Tabla 4.3. Variables de la base de datos de cantidades. Fuente: Elaboración propia.

Variable	Etiqueta	Valores	Etiqueta de valores	Tipo de variable	Formato
macrozona	Macrozona	1	Norte	Discreta	Numérica
		2	Gran Santiago		
		3	Centro		
		4	Sur		
folio_v	Identificador de vivienda	Rango: 1-14961		Discreta	Numérica
		Válidos			
cantidad	Cantidad o número de bienes o servicios adquiridos	Rango: 4-418045		Continua	Numérica
		Válidos			
unidad_medida	Unidad de medida de los bienes o servicios adquiridos			Discreta	Cadena
gasto	Gasto total del hogar por producto	Rango: 62 - 181571		Continua	Numérica
glosa_ccif	Glosa teórica CCIF IX EPF		Válidos	Nominal	Cadena
glosa_establecimiento	Glosa teórica de los establecimientos			Nominal	Cadena

4.3 Preprocesamiento de los datos

El preprocesamiento tuvo como objetivo transformar dos bases relacionales en una matriz analítica única. Primero, se normalizó la lectura de archivos considerando el separador ; y el separador de miles . para evitar que variables numéricas quedaran como cadenas de texto. Segundo, se identificaron códigos especiales de no respuesta, no sabe o no aplica presentes en variables sociodemográficas, particularmente valores negativos como -77, -88

y -99. En variables como edad y escolaridad, dichos valores fueron excluidos del cálculo de promedios.

En la **base de personas**, las variables individuales se agregaron a nivel hogar. Se calculó edad promedio, mediana, máximo y mínimo de edad; proporción de mujeres; proporción de menores de 18 años; proporción de mayores de 65 años; escolaridad promedio; ingreso disponible del hogar; y características del sustentador principal. En la **base de cantidades**, se agregaron el gasto total, cantidad total, número de registros, número de productos distintos, número de establecimientos, gasto promedio por producto, gasto mediano y gasto máximo.

También se construyeron variables específicas para la industria de alimentos y bebidas. Se identificaron registros de bebidas mediante palabras clave en glosa_ccif, tales como bebida, agua, jugo, néctar, café, té, leche, cerveza, vino y alcohol. Luego se calculó gasto_bebidas, cantidad_bebidas, share_bebidas y gasto_alimentos. Adicionalmente, se crearon variables por categorías amplias: pan y cereales, lácteos, carnes, frutas y verduras, azúcar/confites y bebidas azucaradas. Finalmente, se creó el indicador share_supermercados para estimar el peso relativo de este canal en el gasto total.

En lo que respecta las variables utilizadas para los modelos de aprendizaje supervisado y no supervisado, se realizó la estandarización de variables numéricas para los modelos que dependen de distancia o margen, especialmente KNN, SVM y K-Means.

Todos los puntos anteriormente escritos se reflejan de manera resumida en la Tabla 4.4.

Tabla 4.4. Pasos de preprocesamiento realizados. Fuente: Elaboración propia.

Paso	Qué se hizo	Resultado
Lectura	Carga CSV con ; y miles .	Datos interpretados correctamente
Limpieza	Conversión numérica y tratamiento de códigos especiales	Variables comparables y sin errores de formato
Agregación personas	Resumen de edad, escolaridad, sexo, ingreso y sustentador principal por hogar	Perfil sociodemográfico del hogar
Agregación cantidades	Suma de gasto/cantidad y conteo de productos/establecimientos	Perfil de consumo del hogar
Creación de Variables	Creación de ratios per cápita, share de bebidas, canales y categorías	Variables explicativas para ML
Base final	Integración por folio_v y conservación de CSE válido	13.316 hogares modelables

4.4 Construcción de los modelos

La construcción de modelos se organizó en dos bloques: modelos supervisados y modelos no supervisados. Los modelos supervisados se utilizaron para **predecir la clasificación socioeconómica del hogar (CSE)** a partir de variables de consumo, ingreso, composición del hogar y canales de compra. La variable objetivo fue cse con tres clases: 1 = bajo, 2 = medio y 3 = alto. Esta elección es consistente con el enfoque de segmentación y predicción, ya que permite evaluar si los patrones observados de gasto y consumo contienen información suficiente para distinguir perfiles socioeconómicos.

Se implementaron tres algoritmos: Regresión Logística, K-Nearest Neighbors (KNN) y Support Vector Machine (SVM). La Regresión Logística se usó como modelo interpretable de referencia; KNN se incorporó porque clasifica hogares según similitud en el espacio de variables; y SVM se utilizó por su capacidad de construir fronteras de decisión robustas en espacios de alta dimensionalidad.

El modelamiento no supervisado se utilizó para **descubrir segmentos de hogares** sin usar la etiqueta CSE como variable objetivo. Para ello se aplicó K-Means con el objetivo de agrupar hogares según similitudes multivariadas en consumo, gasto, ingreso y composición del hogar.

4.4.1 Modelos supervisados

Los modelos supervisados se entrenaron con pares (X, y) , donde X representa el conjunto de variables explicativas e y la clase real del hogar. Las variables de entrada incluyeron macrozona, tamaño del hogar, edad promedio, composición etaria, escolaridad, ingreso del hogar, ingreso per cápita, gasto total, gasto per cápita, cantidad total, número de productos, número de establecimientos, participación del gasto en bebidas, participación del gasto en supermercados y gastos por categorías de alimentos y bebidas.

La base fue dividida en entrenamiento y prueba con una partición 75% - 25%, utilizando muestreo estratificado por cse para conservar la proporción de hogares de nivel bajo, medio y alto en ambos subconjuntos. Las variables numéricas fueron imputadas por mediana y estandarizadas con StandardScaler. Las variables categóricas, como macrozona y sexo del sustentador principal, fueron imputadas por moda y codificadas mediante One-Hot Encoding.

4.4.1.1 Regresión logística multiclase

La regresión logística se utilizó como modelo base interpretable para clasificación. Aunque su nombre incluye la palabra regresión, en este estudio cumple una función de clasificación, porque estima la probabilidad de pertenencia de cada hogar a una clase socioeconómica. En el caso binario, la regresión logística modela la probabilidad de ocurrencia de una clase mediante la función logística:

$$P(y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

donde $P(y = 1 | X)$ es la probabilidad estimada de que el hogar pertenezca a una clase, β_0 es el intercepto, β_j son los coeficientes del modelo y x_j son las variables explicativas. La

función logística transforma una combinación lineal de variables en un valor entre 0 y 1, interpretable como probabilidad.

Para el caso multiclase utilizado en esta memoria, la lógica se extiende a una función softmax. Para cada clase k , el modelo calcula:

$$P(y = k | X) = \frac{e^{(\beta_{k0} + \beta_{k1}x_1 + \dots + \beta_{kp}x_p)}}{\sum_j e^{(\beta_{j0} + \beta_{j1}x_1 + \dots + \beta_{jp}x_p)}}$$

La predicción final corresponde a la clase con mayor probabilidad estimada. Este modelo es útil porque permite identificar relaciones lineales entre variables de ingreso, gasto y composición del hogar con la clasificación socioeconómica. Además, entrega probabilidades comparables entre clases, lo que permite calcular Curva ROC y ROC-AUC en enfoque comparativo.

Paso a paso aplicado: primero se imputaron valores faltantes; luego se estandarizaron las variables numéricas; posteriormente se codificaron las variables categóricas; después se entrenó LogisticRegression con `class_weight="balanced"` para reducir el efecto del desbalance entre clases; finalmente se predijeron las clases sobre el conjunto de prueba y se calcularon métricas de desempeño.

4.4.1.2 *K-Nearest Neighbors (KNN)*

KNN es un algoritmo supervisado basado en distancia. No estima coeficientes ni construye una función paramétrica explícita; en cambio, memoriza el conjunto de entrenamiento y clasifica un nuevo hogar según las clases de sus k vecinos más cercanos. En este estudio se utilizó $k = 11$ y ponderación por distancia, de modo que los vecinos más cercanos influyen más que los vecinos lejanos.

La distancia euclidiana entre dos hogares x_i y x_j se calcula como:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^p (x_{im} - x_{jm})^2}$$

La clase predicha para un hogar x se obtiene por mayoría ponderada entre sus vecinos:

$$\hat{y} = \underset{c}{\operatorname{argmax}} \sum_{i \in N_k(x)} w_i \cdot I(y_i = c)$$

$$\text{con } w_i = \frac{1}{d(x, x_i) + \varepsilon}$$

donde $N_k(x)$ es el conjunto de k vecinos más cercanos, $I(y_i = c)$ es una función indicadora que vale 1 si el vecino pertenece a la clase c , w_i es el peso por distancia y ε evita divisiones por cero. KNN es útil en esta tesis porque permite capturar relaciones no lineales entre variables de consumo e ingreso, especialmente cuando hogares con comportamientos similares tienden a compartir una clasificación socioeconómica.

Paso a paso aplicado: se estandarizaron todas las variables numéricas, ya que KNN es sensible a la escala; se codificaron las variables categóricas; se definió $k = 11$ para balancear ruido y capacidad predictiva; se entrenó el pipeline; luego, para cada hogar del conjunto de prueba, el algoritmo buscó los 11 hogares más similares del conjunto de entrenamiento y asignó la clase por voto ponderado.

4.4.1.3 Support Vector Machine (SVM)

SVM es un modelo supervisado que busca separar clases mediante un hiperplano óptimo. En clasificación binaria, el objetivo es encontrar el hiperplano $w^T x + b = 0$ que

maximice el margen entre clases. El margen corresponde a la distancia entre el hiperplano y los puntos más cercanos de cada clase, conocidos como vectores de soporte.

La función de decisión se expresa como:

$$f(x) = w^T x + b$$

La clase se asigna según el signo de $f(x)$. En su forma de margen suave, el problema de optimización es:

$$\min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$$

sujeto a: $y_i(w^T x_i + b) \geq 1 - \xi_i$; $\xi_i \geq 0$

donde $\|w\|^2$ controla la complejidad del hiperplano, C regula la penalización por errores de clasificación y ξ_i son variables de holgura que permiten errores cuando los datos no son perfectamente separables. Para el problema multiclase, SVM utiliza una estrategia de comparación entre clases, como One-vs-Rest, donde se entrena un clasificador por clase contra las demás.

SVM es útil en este estudio porque maximiza la separación entre grupos socioeconómicos en un espacio de variables estandarizadas. Al ser un modelo de margen, suele comportarse bien cuando existen muchas variables explicativas y las diferencias entre clases son sutiles. En esta memoria se utilizó LinearSVC por eficiencia computacional y `class_weight="balanced"` para corregir desbalance de clases.

Paso a paso aplicado: se prepararon las variables mediante imputación, escalamiento y codificación; se entrenó LinearSVC sobre el conjunto de entrenamiento; el modelo estimó hiperplanos de separación entre las clases Bajo, Medio y Alto; posteriormente se evaluó sobre la muestra de prueba mediante matriz de confusión y métricas macro-promediadas.

4.4.2 Modelos no supervisados

4.4.2.1 K-Means

El modelo no supervisado utilizado fue K-Means, cuyo objetivo es dividir los hogares en k grupos o clústers, de manera que los hogares dentro de un mismo clúster sean lo más parecidos posible entre sí y, al mismo tiempo, distintos de los hogares pertenecientes a otros clústers. En este caso no se utilizó la variable cse como etiqueta de entrenamiento; solo se utilizaron variables de gasto, ingreso, composición del hogar y patrones de consumo.

Formalmente, K-Means busca minimizar la suma de cuadrados intra-clúster:

$$J = \sum_{k=1}^K \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

donde C_k es el clúster k , x_i es un hogar representado por sus variables estandarizadas, μ_k es el centroide del clúster k y $||x_i - \mu_k||^2$ es la distancia cuadrática entre el hogar y su centroide. El objetivo es encontrar centroides que reduzcan la dispersión interna de cada grupo.

El centroide de cada clúster se calcula como el promedio vectorial de las observaciones asignadas:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

K-Means es útil para esta investigación porque permite construir segmentos de hogares basados en comportamiento real de consumo, sin imponer previamente grupos tradicionales. Esto permite reconocer perfiles de gasto intensivo, hogares con alta participación de bebidas, consumidores diversificados o segmentos con mayor concentración en supermercados.

Paso a paso aplicado en la tesis:

- Selección de variables de segmentación: gasto total, gasto per cápita, gasto en bebidas, share de bebidas, número de productos, share supermercados, ingreso per cápita, número de personas, edad promedio, escolaridad promedio y gasto por categorías de alimentos.
- Imputación de valores faltantes mediante mediana para evitar pérdida de hogares.
- Estandarización de variables para que ingreso o gasto no dominen artificialmente la distancia euclidiana.
- Entrenamiento de K-Means para valores de k entre 2 y 8.
- Cálculo de inercia para aplicar el método del codo.
- Cálculo de Average Silhouette para analizar cohesión y separación de los clústers.
- Selección final de $k = 4$, por equilibrio entre interpretabilidad, reducción de inercia y calidad de separación.
- Perfilamiento de clústers mediante promedios de gasto, ingreso, composición del hogar, educación, participación de bebidas y uso de supermercados.

4.5 Evaluación de los modelos

4.5.1 Modelos supervisados

La evaluación de los modelos supervisados se basó en la **matriz de confusión** y en métricas derivadas de ella. La matriz de confusión compara las clases reales con las clases predichas, permitiendo observar aciertos y errores por clase. En clasificación binaria, sus componentes son:

- TP (True Positive): casos positivos correctamente clasificados como positivos.
- TN (True Negative): casos negativos correctamente clasificados como negativos.
- FP (False Positive): casos negativos clasificados erróneamente como positivos.
- FN (False Negative): casos positivos clasificados erróneamente como negativos.

En clasificación multiclase, como en esta tesis, se evalúa cada clase con enfoque One-vs-Rest: una clase se considera positiva y las demás se agrupan como negativas. Luego se calculan las métricas por clase y se promedian. Se utilizó promedio macro, porque entrega el mismo peso a cada clase, independiente del número de hogares que tenga.

A continuación se describen los indicadores de desempeños que se obtienen de la matriz de confusión:

Accuracy

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Mide la proporción total de predicciones correctas. Es útil para conocer el desempeño global, pero puede ser optimista cuando existen clases desbalanceadas.

Recall o Sensibilidad

$$Recall = \frac{TP}{(TP + FN)}$$

Mide la capacidad del modelo para detectar correctamente los casos reales de una clase. Es útil cuando interesa reducir falsos negativos.

Specificity o Especificidad

$$Specificity = \frac{TN}{(TN + FP)}$$

Mide la capacidad del modelo para reconocer correctamente los casos que no pertenecen a una clase. Es útil para controlar falsos positivos.

Precision

$$Precision = \frac{TP}{(TP + FP)}$$

Mide qué proporción de los casos predichos como una clase realmente pertenece a esa clase. Es útil cuando se busca que las asignaciones positivas sean confiables.

F-score o F1-score

$$F - score = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)}$$

Resume Precision y Recall en una sola métrica. Es útil cuando se requiere equilibrio entre detectar casos reales y evitar clasificaciones incorrectas.

Curva ROC

$$TPR = \frac{TP}{(TP + FN)}; \quad FPR = \frac{FP}{(FP + TN)}$$

Representa la relación entre tasa de verdaderos positivos y tasa de falsos positivos para distintos umbrales de decisión. En multiclase se aplica con One-vs-Rest.

ROC-AUC

$$AUC = \text{área bajo la curva ROC}$$

Mide la capacidad del modelo para ordenar correctamente observaciones de una clase frente al resto. Un valor cercano a 1 indica alta capacidad discriminativa; un valor cercano a 0,5 equivale a clasificación aleatoria.

4.5.2 Modelos no supervisados

En modelos no supervisados no existe una etiqueta real contra la cual comparar la predicción. Por ello, la evaluación se realiza con métricas internas que analizan la estructura de los grupos encontrados. En esta investigación se aplicaron el método del codo y el Average Silhouette.

4.5.2.1 Método del codo (Elbow)

El método del codo utiliza la inercia o suma de cuadrados intra-clúster para distintos valores de k. La inercia se calcula como:

$$Inercia(k) = \sum_{k=1}^K \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

A medida que aumenta k, la inercia siempre disminuye, porque los grupos se vuelven más pequeños. Sin embargo, llega un punto en que agregar más clústers reduce muy poco la inercia. Ese punto de cambio se conoce como codo. Es útil porque ayuda a seleccionar un número de clústers que entregue buena compactación sin sobresegmentar los hogares.

4.5.2.2 Average Silhouette

El coeficiente Silhouette evalúa simultáneamente la cohesión interna del clúster y la separación frente a otros clústers. Para cada observación i se calcula:

$$s(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}}$$

donde $a(i)$ es la distancia promedio de la observación i respecto de los demás hogares de su mismo clúster, y $b(i)$ es la menor distancia promedio de i respecto de los hogares del clúster vecino más cercano. El valor de $s(i)$ se ubica entre -1 y 1.

- Valores cercanos a 1 indican que el hogar está bien asignado: es parecido a su clúster y distinto de los otros.
- Valores cercanos a 0 indican solapamiento entre clústers.
- Valores negativos sugieren que el hogar podría estar mejor asignado a otro clúster.

El Average Silhouette corresponde al promedio de $s(i)$ en todos los hogares, se calcula:

$$Average\ Silhouette = \frac{1}{n} \sum_{i=1}^n s(i)$$

Es útil para comparar distintos valores de k desde la perspectiva de la calidad de separación. En esta tesis se usó como complemento al método del codo, priorizando una solución interpretable y con coherencia estratégica.

4.6 Herramientas utilizadas

El procesamiento se realizó con Python. Se utilizaron pandas y numpy para limpieza e integración; matplotlib para visualización; scikit-learn para pipelines, preprocesamiento, modelos supervisados, K-Means y métricas. La estrategia permitió trazabilidad completa desde la carga de datos hasta la generación de tablas, gráficos y anexos reproducibles.

5. Resultados

5.1 Análisis descriptivo de los datos

La base integrada final quedó conformada por 13.316 hogares. El gasto total promedio por hogar en alimentos y bebidas fue de 30.393 CLP. con una mediana de 21.825 CLP. Esta diferencia entre promedio y mediana evidencia asimetría positiva: existe un grupo de hogares con gastos altos que eleva el promedio general. El ingreso disponible promedio del hogar fue de 1.357.797 CLP. mientras que el ingreso per cápita promedio alcanzó 509.787 CLP.

Desde una perspectiva de negocio, estos resultados muestran que el consumo de alimentos y bebidas no depende únicamente del ingreso, sino también de variables de tamaño del hogar, canal de compra, número de categorías adquiridas y composición del gasto. Por ello, la segmentación mediante Machine Learning es pertinente: permite combinar variables de poder adquisitivo con patrones efectivos de compra.

Tabla 5.1: Estadísticos descriptivos de variables principales de la base integrada. Fuente: Elaboración propia.

Variable	N	Media	Desv. Est.	Mín.	P25	Mediana	P75	P90	Máx.
gasto_total	13.316	30.392,84	30.293,54	207,00	10.142,00	21.825,00	40.402,50	66.635,00	485.395,00
gasto_per_capita	13.316	11.341,30	11.757,53	41,30	3.914,12	8.052,75	14.836,62	24.392,25	172.504,00
ing_disp_hog_hd	13.235	1.357.797,41	1.498.779,42	-500.000,00	548.950,00	939.637,00	1.600.618,00	2.746.013,20	34.590.449,00
ingreso_per_capita	13.235	509.787,29	609.077,04	-315.000,00	206.250,00	341.258,75	584.562,50	1.028.574,68	17.001.524,00
gasto_bebidas	13.316	13.798,14	18.600,08	0,00	2.214,00	7.972,00	18.578,00	33.612,00	467.716,00
share_bebidas	13.316	0,44	0,34	0,00	0,14	0,41	0,70	1,00	1,00
cantidad_total	13.316	26.257,15	33.887,60	17,00	6.749,75	16.163,00	32.991,25	58.786,50	450.056,00
npersonas	13.316	3,04	1,52	1,00	2,00	3,00	4,00	5,00	16,00
edad_prom	13.316	41,31	17,64	10,00	27,00	37,00	54,00	68,50	98,00
edue_prom	13.316	10,71	3,60	0,00	8,00	10,50	13,00	15,69	24,00
n_productos	13.316	5,38	3,49	1,00	3,00	5,00	8,00	10,00	22,00
n_establecimientos	13.316	1,98	0,93	1,00	1,00	2,00	3,00	3,00	6,00

El gasto promedio más alto se observa en Gran Santiago, seguido por Norte, Sur y Centro. Esta diferencia es consistente con una mayor concentración urbana, mayor disponibilidad de canales modernos y mayor ingreso promedio relativo en zonas metropolitanas. Por clasificación socioeconómica, el gasto promedio aumenta desde CSE Bajo hacia CSE Alto; sin embargo, el salto más relevante se aprecia en ingresos, lo que sugiere que el gasto alimentario posee una base necesaria común y que la diferenciación por ingreso se expresa con mayor claridad en gasto per cápita, diversidad de productos y canal de compra.

Tabla 5.2: Gasto e ingreso promedio por macrozona, incluyendo cantidad de hogares.

Fuente: Elaboración propia.

Macrozona	Hogares	Gasto prom. hogar	Ingreso prom. hogar	Gasto prom. per cápita	Ingreso prom. per cápita	Personas prom.
Norte	2376	32.027	1.216.439	11.377	426.286	3,28
Gran Santiago	5293	35.182	1.602.415	12.941	597.625	3,08
Centro	3574	25.958	1.151.043	9.911	443.546	2,93
Sur	2073	23.938	1.249.496	9.682	494.663	2,83

Tabla 5.3: Gasto e ingreso promedio por clasificación socioeconómica, incluyendo cantidad de hogares. Fuente: Elaboración propia.

CSE	Hogares	Gasto prom. hogar	Ingreso prom. hogar	Gasto prom. per cápita	Ingreso prom. per cápita	Personas prom.
Bajo	3451	27.641	835.877	9.531	280.339	3,28
Medio	5910	29.411	1.161.903	10.922	423.917	3,03
Alto	3955	34.262	2.104.390	13.547	837.634	2,84

Figura 5.1: Gasto promedio por macrozona. Fuente: (Elaboración propia)

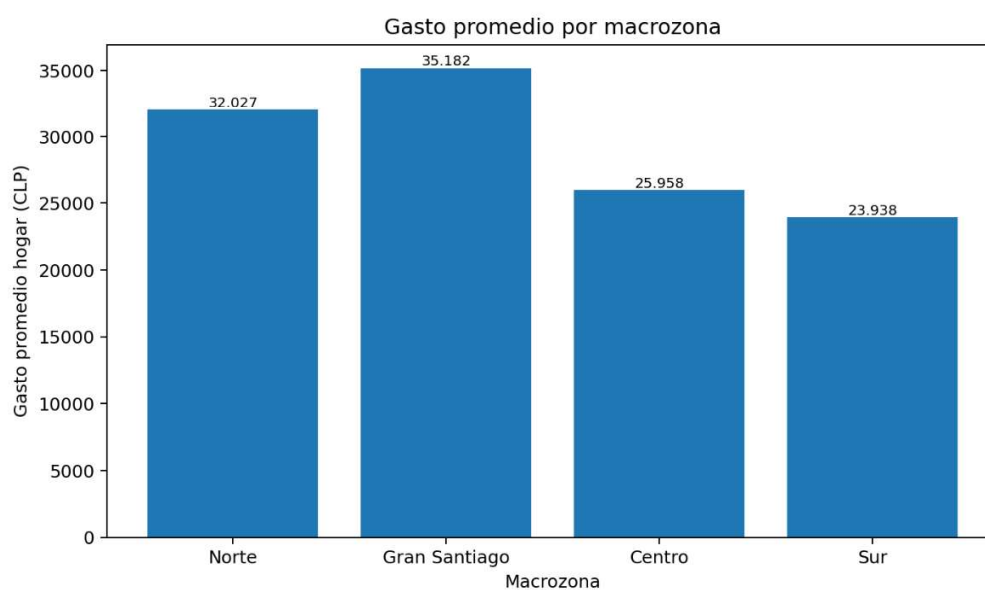


Figura 5.2: *Ingreso promedio por macrozona. Fuente: (Elaboración propia)*

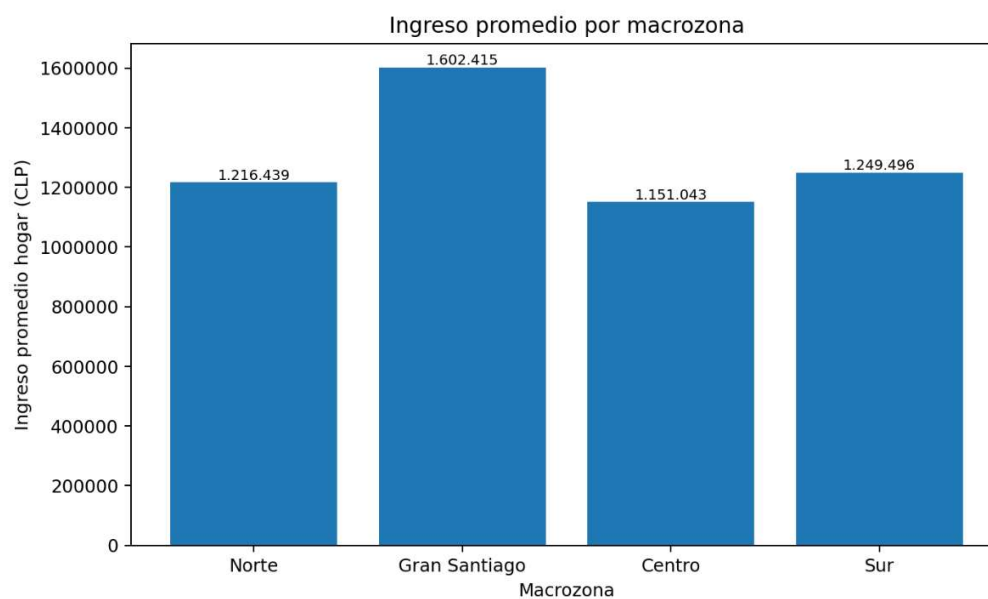


Figura 5.3: *Gasto promedio por clasificación socioeconómica. Fuente: (Elaboración propia)*

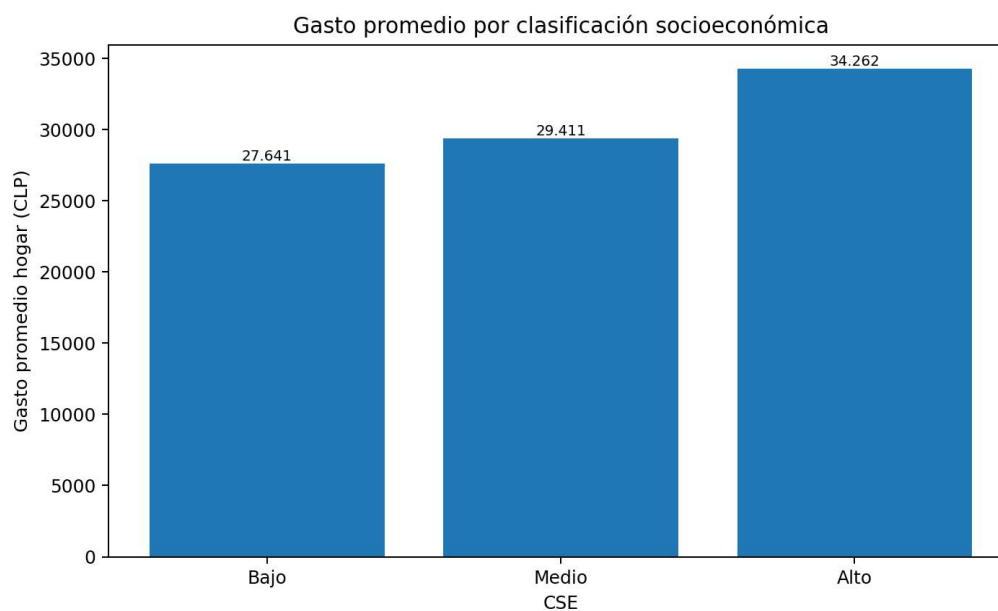
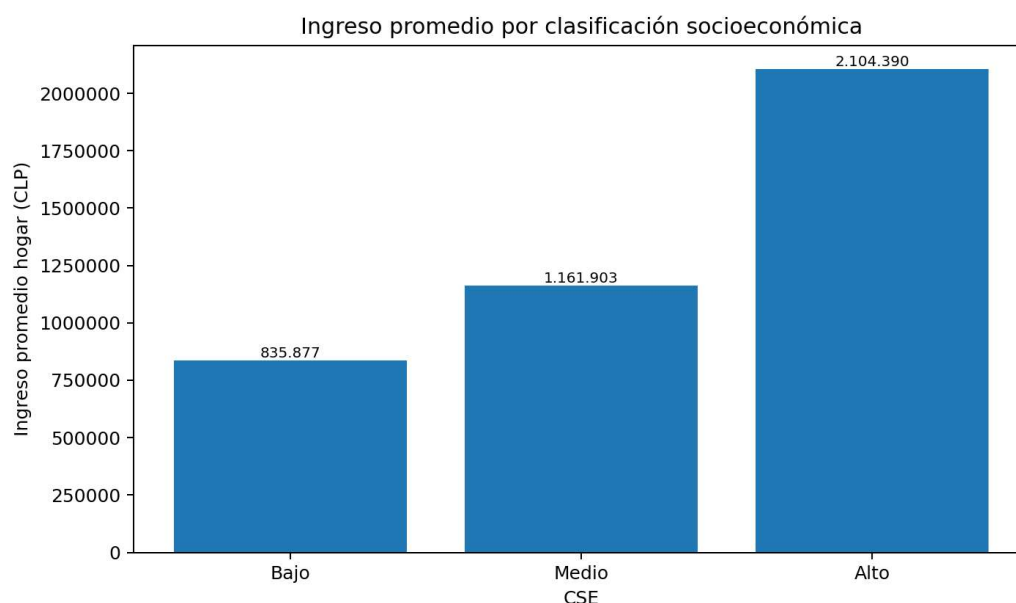


Figura 5.4: Ingreso promedio por clasificación socioeconómica. Fuente: (Elaboración propia)



El análisis por establecimiento muestra que supermercados, ferias libres y almacenes tradicionales concentran una parte importante del gasto total. La inclusión de la cantidad de hogares permite distinguir entre establecimientos de alto gasto total por masividad y establecimientos con menor alcance pero tickets potencialmente superiores.

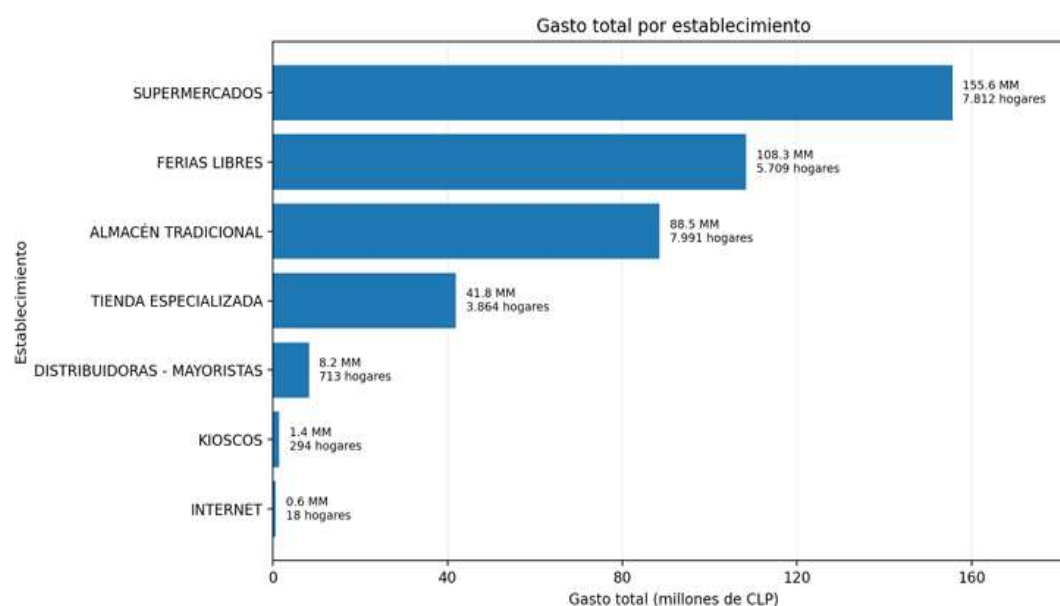
Tabla 5.4: Principales establecimientos por gasto total, registros y cantidad de hogares.

Fuente: Elaboración propia.

Establecimiento	Gasto total	Registros	Hogares
SUPERMERCADOS	155.649.142	37.999	7.812
FERIAS LIBRES	108.342.658	37.124	5.709
ALMACÉN TRADICIONAL	88.540.730	34.127	7.991
TIENDA ESPECIALIZADA	41.841.188	11.460	3.864

DISTRIBUIDORAS MAYORISTAS	- 8.245.415	1.383	713
KIOSCOS	1.442.288	602	294
INTERNET	649.624	28	18

Figura 5.5: *Gasto total por establecimiento de compra. Fuente: (Elaboración propia)*



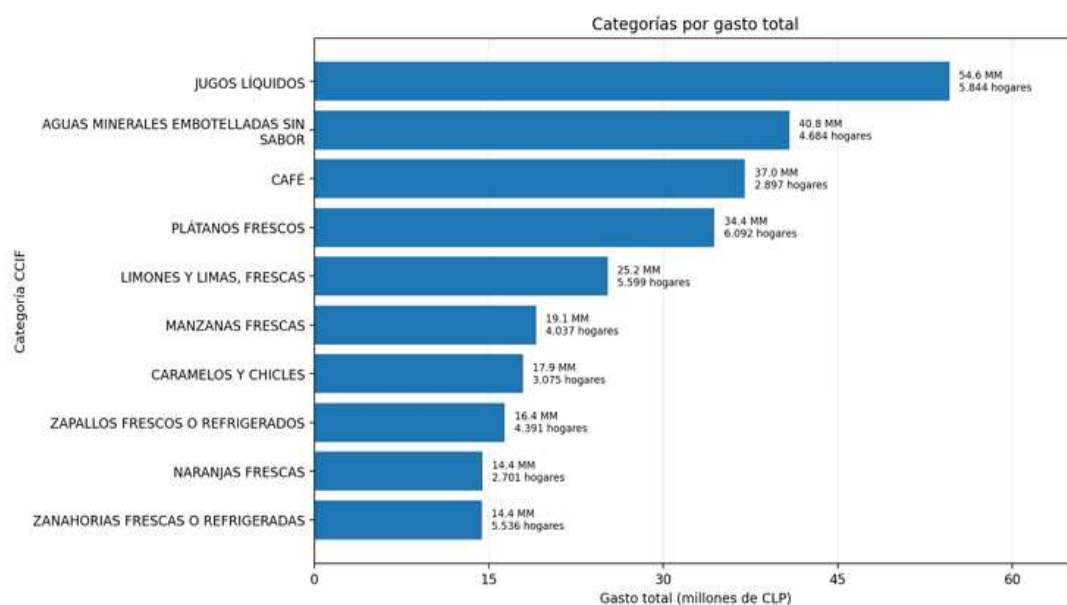
Las categorías con mayor gasto total incluyen jugos líquidos, aguas minerales sin sabor, café y frutas frescas como plátanos, limones y limas. Este resultado es importante para la industria de bebidas, porque evidencia que el consumo líquido no se reduce a bebidas gaseosas, sino que incluye agua, jugos, café y otros productos asociados a conveniencia, hidratación y hábitos cotidianos.

Tabla 5.5: Principales categorías por gasto total, registros y cantidad de hogares. Fuente:

Elaboración propia.

Categoría CCIF	Gasto total	Registros	Hogares
JUGOS LÍQUIDOS	54.616.903	14.949	5.844
AGUAS MINERALES	40.839.611	10.217	4.684
EMBOTELLADAS SIN SABOR	37.019.756	4.304	2.897
CAFÉ			
PLÁTANOS FRESCOS	34.395.785	10.424	6.092
LIMONES Y LIMAS, FRESCAS	25.223.970	9.378	5.599
MANZANAS FRESCAS	19.067.910	6.229	4.037
CARAMELOS Y CHICLES	17.922.064	5.545	3.075
ZAPALLOS FRESCOS O REFRIGERADOS	16.378.119	7.488	4.391
NARANJAS FRESCAS	14.447.286	3.970	2.701
ZANAHORIAS FRESCAS O REFRIGERADAS	14.413.302	8.863	5.536

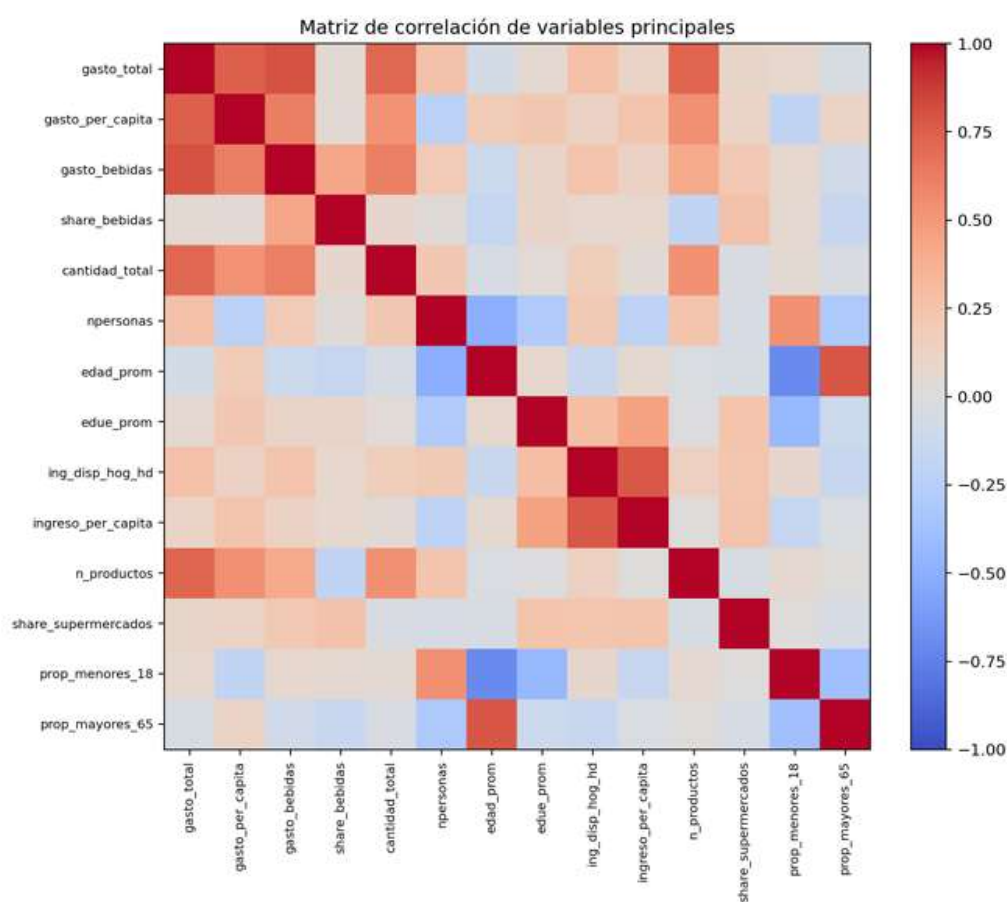
Figura 5.6: Categorías por gasto total. Fuente: (Elaboración propia)



5.1.1 Matriz de correlación de variables principales

La matriz de correlación de variables principales evidenció relaciones esperables y útiles para el modelamiento: el gasto total se asocia positivamente con la cantidad total, el número de productos y el gasto en bebidas; el ingreso per cápita se relaciona con el nivel de escolaridad promedio; y el share de supermercados permite diferenciar patrones de canal de compra. Estas correlaciones justifican el uso de modelos multivariados, ya que ninguna variable aislada explica por completo la segmentación de hogares.

Figura 5.7: Matriz de correlación de variables principales. Fuente: (Elaboración propia)



5.2 Resultados de modelos supervisados

Los modelos supervisados fueron entrenados para **predecir la CSE de los hogares** a partir de variables de consumo, composición del hogar, ingreso, productos y canales. El conjunto de prueba consideró 3.329 hogares. Las métricas fueron calculadas en promedio macro para asignar el mismo peso a cada clase, evitando que la clase mayoritaria domine la evaluación.

Los resultados muestran desempeños moderados. SVM obtuvo el mayor Accuracy (56,32%) y el mayor F1-score (56,67%), mientras que Regresión Logística obtuvo el mejor

ROC-AUC (76,23%), ligeramente por sobre SVM (74,97%). KNN presentó un desempeño inferior, con Accuracy de 52,24% y F1-score de 50,99%. En términos prácticos, esto indica que los patrones de consumo permiten distinguir parcialmente los niveles socioeconómicos, pero no reemplazan mediciones directas de ingreso, educación y composición socioeconómica.

La Regresión Logística y SVM muestran resultados similares porque ambos son modelos de frontera relativamente global. KNN fue más sensible al ruido y a la cercanía entre hogares con patrones de consumo similares pero CSE distintos. Esto es esperable en datos de consumo familiar, donde dos hogares pueden comprar productos similares por razones distintas: tamaño del hogar, preferencias, territorio o canal de compra.

Tabla 5.6: Métricas comparativas de modelos supervisados. Fuente: Elaboración propia.

Modelo	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC
Regresión Logística	56,23%	56,71%	59,86%	78,37%	56,56%	76,23%
KNN	52,24%	52,45%	50,33%	74,31%	50,99%	69,95%
SVM	56,32%	56,18%	58,81%	78,01%	56,67%	74,97%

La matriz de confusión de Regresión Logística muestra una buena identificación de los hogares CSE alto: 716 aciertos sobre 989 casos reales de CSE alto. Sin embargo, existe confusión entre CSE bajo y medio, lo que refleja que estos segmentos comparten patrones de consumo más próximos. En SVM se observa un comportamiento parecido, con 711 aciertos para CSE alto y mejores aciertos en CSE medio que KNN.

Desde la perspectiva de negocio, el error entre bajo y medio es menos crítico que confundir sistemáticamente hogares altos con bajos, porque ambos primeros grupos pueden

compartir estrategias comerciales de precio y formatos de cercanía. La mayor separabilidad del CSE alto sugiere que las variables de ingreso, educación, canal supermercado y gasto per cápita permiten identificar segmentos de mayor poder adquisitivo con mejor precisión relativa.

Figura 5.8: *Matriz de confusión de Regresión Logística. Fuente: (Elaboración propia)*

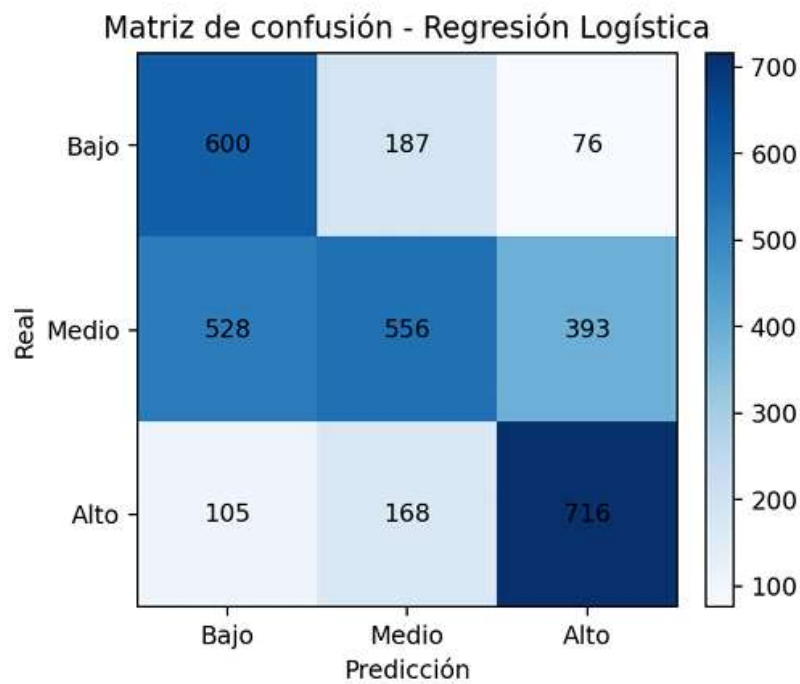


Figura 5.9: *Matriz de confusión - KNN. Fuente: (Elaboración propia)*

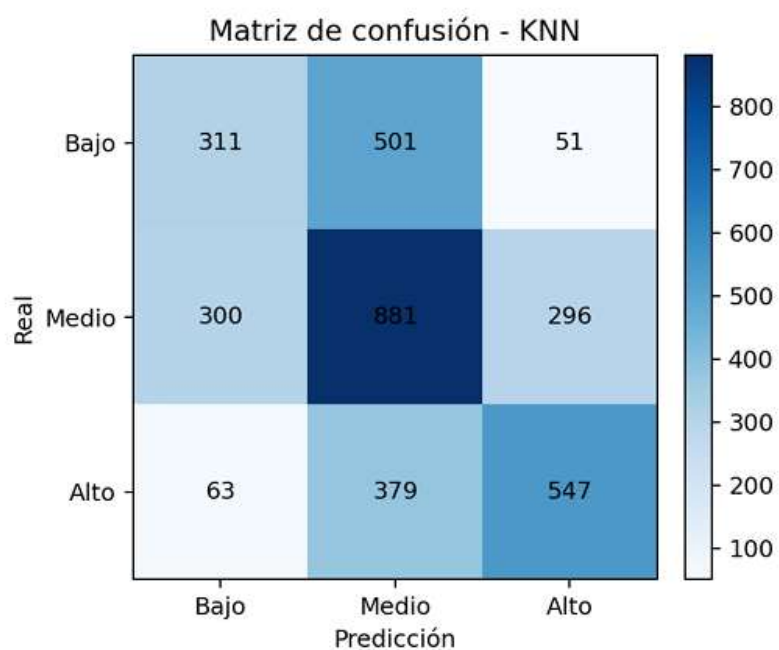


Figura 5.10: *Matriz de confusión - SVM. Fuente: (Elaboración propia)*

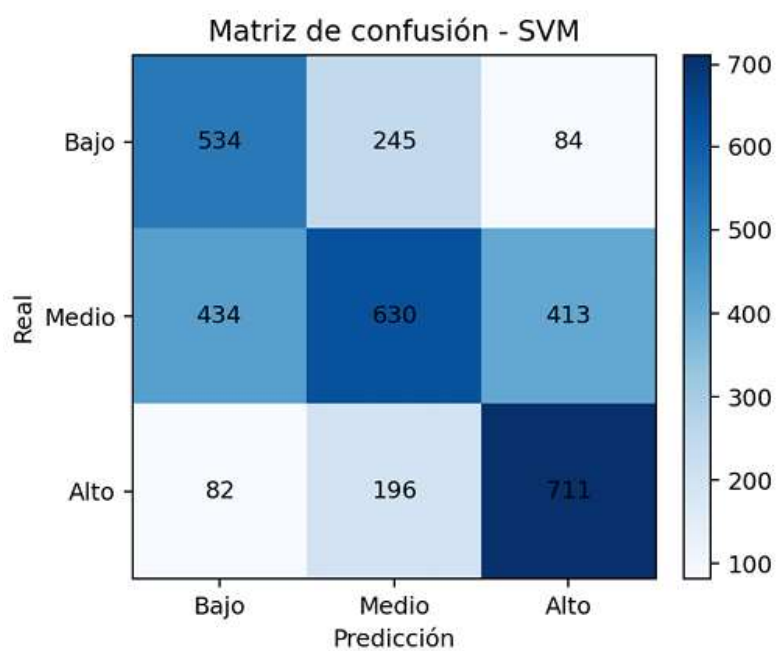
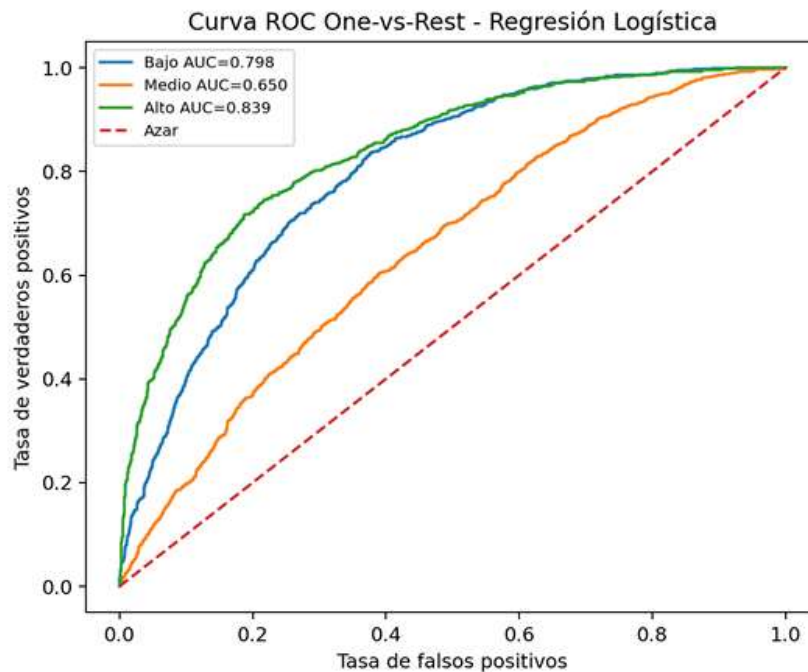


Figura 5.11: Curva ROC One-vs-Rest del modelo con mayor ROC-AUC macro. Fuente: (Elaboración propia)



5.3 Resultados de modelos no supervisados

Para los modelos no supervisados se evaluaron soluciones de $k=2$ a $k=8$. Para el método del codo la inercia disminuyó desde 155.683,8 en $k=2$ hasta 118.910,6 en $k=4$. Posteriormente, la reducción continúa pero con ganancias marginales menores. Este patrón respalda la selección de $k=4$ como un punto de equilibrio entre explicación de variabilidad y simplicidad interpretativa.

El método Average Silhouette alcanzó su máximo en $k=2$ (0,202) y luego disminuyó gradualmente hasta 0,179 en $k=4$. Aunque $k=2$ separa mejor los datos desde una perspectiva puramente geométrica, resulta demasiado general para una tesis orientada a segmentación de clientes, ya que dividiría el mercado en solo dos grupos y limitaría la utilidad de

recomendaciones comerciales. Por ello se seleccionó $k=4$, considerando simultáneamente el codo, la silueta y la interpretabilidad de perfiles.

Tabla 5.7: Inercia y Average Silhouette para selección de k . Fuente: Elaboración propia.

k	Inercia	silhouette
2,0000	155.683,79	0,2016
3,0000	134.724,00	0,1912
4,0000	118.910,60	0,1788
5,0000	112.930,07	0,1493
6,0000	107.261,89	0,1360
7,0000	102.558,67	0,1428
8,0000	96.815,53	0,1361

Figura 5.12: Método del codo para K-Means. Fuente: (Elaboración propia)

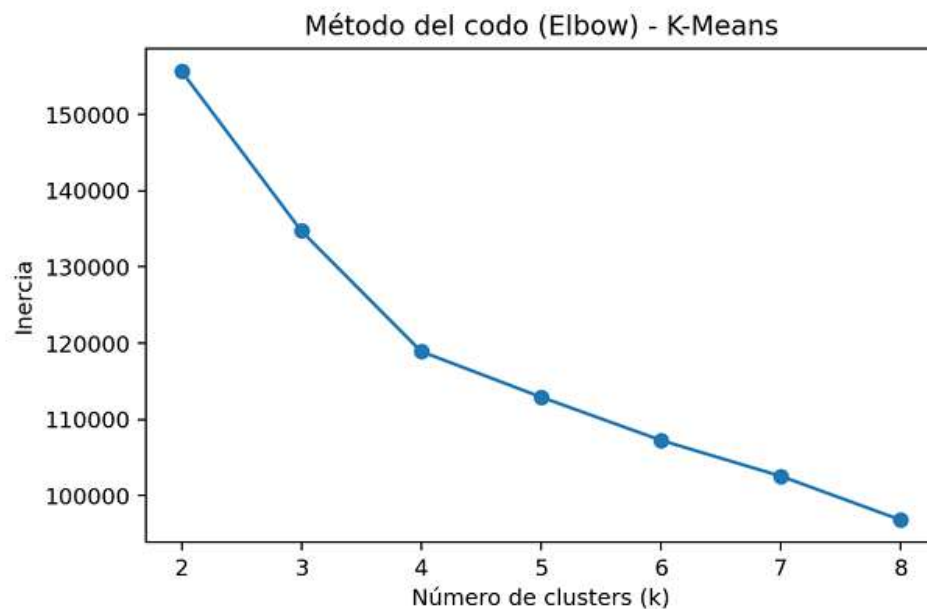
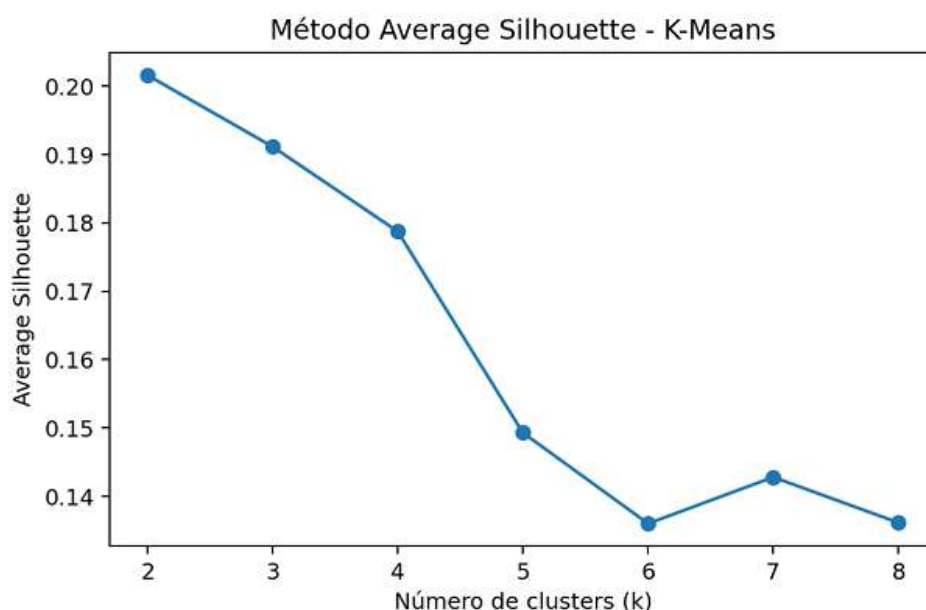


Figura 5.13: Método Average Silhouette para K-Means. Fuente: (Elaboración propia)



5.4 Análisis de perfiles de clústers

Clúster 1 - Familias jóvenes de gasto moderado: agrupa 4.481 hogares, con 4,05 personas promedio, edad promedio de 27,65 años y escolaridad promedio de 8,67 años. Su gasto total promedio fue \$22.058 y su gasto per cápita \$5.823. Es un grupo de hogares numerosos, relativamente jóvenes y con restricción presupuestaria por persona. Para la industria, este segmento es sensible a formatos familiares, promociones por volumen y disponibilidad en canales de cercanía.

Clúster 2 - Hogares mayores de consumo conservador: reúne 2.516 hogares, con 1,91 personas promedio y edad promedio de 68,17 años. Su gasto total promedio fue \$21.431, con gasto per cápita de \$12.284 y baja participación de bebidas (28%). Corresponde a hogares pequeños y mayores, con consumo más estable y menos intensivo en variedad. Las

estrategias recomendadas son formatos pequeños, productos tradicionales, información clara y canales de compra presenciales.

Clúster 3 - Compradores selectivos de mayor ingreso: incluye 4.044 hogares, con ingreso per cápita promedio de \$724.165, escolaridad promedio de 13,42 años y share de supermercados de 45%. Su gasto total promedio fue menor (\$17.505), pero su ingreso y escolaridad sugieren hogares con mayor capacidad de elección y compras más selectivas. Este segmento puede responder a productos saludables, premium, funcionales o con atributos de sostenibilidad.

Clúster 4 - Consumidores intensivos multicanal: agrupa 2.275 hogares y presenta el gasto total promedio más alto (\$79.630), gasto per cápita de \$25.351, gasto en bebidas de \$40.878 y 9,65 productos distintos en promedio. Es el segmento de mayor valor comercial para acciones de fidelización, promociones cruzadas, lanzamientos y planificación de demanda. Su alto gasto y variedad de productos lo hacen prioritario para estrategias de surtido y disponibilidad.

Figura 5.14: *Visualización PCA de clústers. Fuente: (Elaboración propia)*

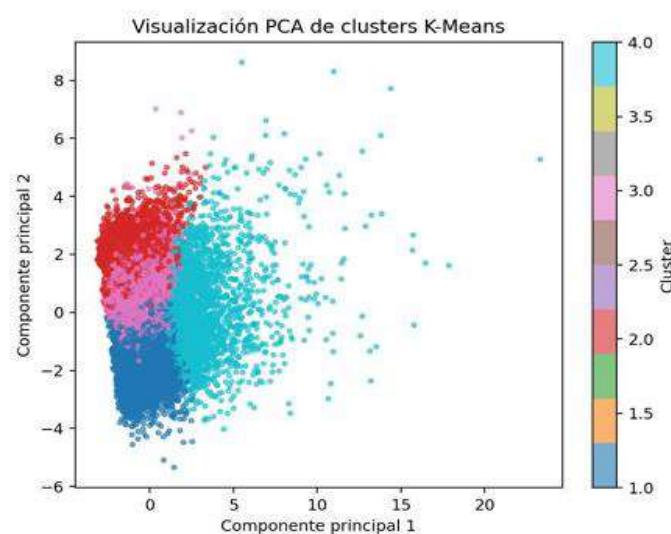


Tabla 5.8: Perfil de clústers obtenidos mediante K-Means. Fuente: Elaboración propia.

Clúster	Hogares	Gasto prom.	Gasto pc	Gasto bebidas	Share bebidas	Ingreso pc	Personas	Edad prom.	Escolari dad	N productos	Share super
1.0	4481.0	22.058	5.823	8.290	40,00%	306.262	4,05	27,65	8,67	4,94	26,00%
2.0	2516.0	21.431	12.284	6.372	28,00%	423.435	1,91	68,17	9,44	5,04	25,00%
3.0	4044.0	17.505	8.989	9.288	53,00%	724.165	2,24	42,03	13,42	3,66	45,00%
4.0	2275.0	79.630	25.351	40.878	53,00%	626.330	3,70	37,21	11,30	9,65	45,00%

5.5 Discusión de hallazgos

Los resultados respaldan la utilidad de Machine Learning para complementar la segmentación tradicional de clientes en alimentos y bebidas. La evidencia muestra que variables de gasto, canales y composición del hogar ayudan a distinguir perfiles, aunque los límites entre grupos no son absolutos. Esta conclusión es importante porque los hogares no se comportan como categorías estáticas: dos hogares con ingresos distintos pueden presentar consumos similares si tienen tamaño familiar parecido, preferencias comunes o acceso a los mismos canales.

El análisis descriptivo evidencia que las bebidas tienen un peso central en la canasta estudiada. El gasto en bebidas presenta la mayor correlación con el gasto total, y varios productos de bebidas aparecen entre los principales ítems por gasto acumulado. Para la industria, esto confirma que la categoría bebidas puede actuar como variable ancla para segmentación, promociones y planificación de surtido.

El análisis por CSE revela diferencias en gasto per cápita, escolaridad, ingreso y canal supermercado. No obstante, el desempeño supervisado moderado muestra que la predicción del CSE no debe depender exclusivamente de consumo observado. En la práctica, los modelos pueden funcionar como una herramienta de apoyo para priorizar acciones comerciales, pero deben complementarse con información transaccional, territorial y actitudinal cuando esté disponible.

Los clústers obtenidos tienen aplicación directa en marketing, pricing y operaciones. El Clúster 4 representa hogares de alto valor y alta variedad, útiles para fidelización y lanzamientos. El Clúster 1 requiere formatos económicos y familiares. El Clúster 2 se orienta a productos tradicionales, porciones pequeñas y comunicación simple. El Clúster 3 puede ser

abordado con propuestas diferenciadas, saludables o premium. Esta segmentación permite pasar desde una visión general del mercado a perfiles accionables para la toma de decisiones.

6. Conclusiones

6.1 Conclusiones generales

La investigación demuestra que es factible aplicar modelos de Machine Learning a datos públicos del INE para analizar patrones de consumo de alimentos y bebidas en hogares chilenos. La integración de la base de personas y la base de cantidades permitió construir una matriz analítica de 13.316 hogares, con variables sociodemográficas, económicas y de comportamiento de compra.

Los modelos supervisados permitieron predecir la clasificación socioeconómica con desempeño moderado. SVM alcanzó el mayor Accuracy (56,32%) y F1 macro (56,67%), mientras que Regresión Logística obtuvo el mayor ROC-AUC macro (76,23%). Estos resultados indican que los patrones de consumo sí contienen señales útiles, especialmente para diferenciar hogares de CSE alto, pero no son suficientes para clasificar perfectamente todos los niveles socioeconómicos.

El modelamiento no supervisado permitió identificar cuatro segmentos de hogares con perfiles diferenciados: familias jóvenes de gasto moderado, hogares mayores de consumo conservador, compradores selectivos de mayor ingreso y consumidores intensivos multicanal. Estos segmentos son más accionables para la industria que una segmentación únicamente demográfica, ya que incorporan gasto, canales, variedad de productos, ingreso, edad y escolaridad.

En términos estratégicos, la principal contribución de la tesis es mostrar que la segmentación basada en datos puede mejorar la toma de decisiones en marketing, surtido,

promociones, planificación de demanda y priorización de clientes. La industria de alimentos y bebidas puede usar modelos de ML no como sustituto del criterio comercial, sino como una herramienta complementaria para orientar decisiones con evidencia cuantitativa.

6.2 Conclusiones específicas

Se logró recolectar, limpiar e integrar las bases de la IX EPF usando folio_v como llave de unión. El resultado fue una base de hogares con variables de ingreso, composición familiar, gasto total, gasto en bebidas, categorías de alimentos, cantidad y canales de compra.

Los modelos supervisados demostraron que la clasificación socioeconómica puede ser parcialmente predicha mediante variables de consumo. El mejor desempeño general fue SVM, aunque Regresión Logística mostró mejor capacidad de separación probabilística medida por ROC-AUC.

K-Means permitió identificar segmentos con sentido comercial. Los clusters no replican simplemente el CSE, sino que combinan intensidad de gasto, edad, tamaño de hogar, escolaridad, ingreso y canal. Esto agrega valor porque permite comprender comportamientos de compra más allá de la clasificación socioeconómica tradicional.

La evaluación comparativa permitió determinar fortalezas y debilidades de cada enfoque. En supervisado, la matriz de confusión evidenció mejor identificación del CSE alto y mayor dificultad entre bajo y medio. En no supervisado, el codo y la silueta mostraron que cuatro clústers entregan un balance razonable entre granularidad y estabilidad.

Referencias

- Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer.
- Agosin, M. (2007). Competitividad de las exportaciones chilenas de alimentos. *Revista CEPAL*, 92, 55–72.
- Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1), 80–98.
- Banco Central de Chile. (2022). *Informe de Cuentas Nacionales 2021*. Santiago: Banco Central.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Bramer, M. (2007). *Principles of data mining*. Springer.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Wadsworth.
- Caba, E. A. (2020). *Metodología para la toma de decisiones mediante la minería de datos*. (Tesis para optar al título de ingeniero civil industrial). Universidad Técnica Federico Santa María. Santiago, Chile.
- Contreras, C. A. (2019). *Modelamiento predictivo mediante aprendizaje supervisado de máquina para datos industriales masivos: caso aplicado a empresa ferroviaria*. (Memoria para optar al título de ingeniero civil industrial). Universidad Técnica Federico Santa María. Santiago, Chile.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

Davenport, T., & Bean, R. (2018). Big companies are embracing analytics, but most still don't have a data-driven culture. *Harvard Business Review*.

Euromonitor International. (2021). *Chile Food and Drink Market Overview*.

FAO. (2022). *The State of Food and Agriculture 2022*. Roma: FAO.

González, A. M. (2019). *Propuesta de guía práctica para la aplicación de tecnologías machine learning en las empresas*. (Tesis de grado). Universidad Técnica Federico Santa María. Santiago, Chile.

Gupta, S., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, 53(8), 1049–1064.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hassan, S. U., Niazi, M. A., & Hussain, M. (2018). Consumer segmentation using clustering techniques for product marketing. *Journal of Retailing and Consumer Services*, 43, 65–73.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Hernández, J., & Ponce, L. (2019). Segmentación de clientes de supermercados mediante clustering. *Revista Mexicana de Administración*, 6(2), 45–63.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Instituto Nacional de Estadísticas (INE). (2023). *Informe Anual de la Industria Manufacturera en Chile*. Santiago: INE.

Kotler, P., & Keller, K. (2016). *Marketing Management* (15th ed.). Pearson.

Kudyba, S. (2014). *Big data, mining, and analytics: Components of strategic decision making*. CRC Press.

Kumar, V., Dixit, A., Javalgi, R. G., Dass, M., & Sridhar, S. (2021). Marketing mix decisions: Machine learning approach. *Journal of the Academy of Marketing Science*, 49(6), 1124–1146.

Laszlo, M., & Mukherjee, S. (2007). A genetic algorithm that exchanges neighboring centers for k-means clustering. *Pattern Recognition Letters*, 28(16), 2359–2366.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.

Martínez-López, F. J., Gázquez-Abad, J. C., & Sousa, C. M. P. (2017). Consumer segmentation in the Spanish wine market. *International Journal of Wine Business Research*, 29(4), 326–345.

Meller, P. (1996). *Un siglo de economía política chilena (1890–1990)*. Santiago: Andrés Bello.

Ministerio de Salud (MINSAL). (2021). *Encuesta Nacional de Consumo de Alimentos (ENCA)*. Santiago: Gobierno de Chile.

- Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86–97.
- Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, 24(12), 1565–1567.
- Ruan, Z., & Siau, K. (2019). Digital marketing in the artificial intelligence and machine learning age. *AMCIS 2019 Proceedings*. Association for Information Systems.
- Runkler, T. A. (2012). *Data analytics: Models and algorithms for intelligent data analysis*. Vieweg+Teubner Verlag.
- Schölkopf, B. (2001). The kernel trick for distances. *Advances in Neural Information Processing Systems*.
- Silva, R., Oliveira, P., & Santos, D. (2020). Predicting energy drink consumption patterns using machine learning. *Revista Brasileira de Marketing*, 19(1), 54–72.
- Soria, P., & Fernández, M. (2021). Segmentación de consumidores de alimentos saludables en Argentina mediante K-Means. *Revista Latinoamericana de Marketing*, 12(3), 112–128.
- Tecoalu, M., Wahyoedi, S., & Kustiawan, E. (2021). The effect of perception of security, ease of transaction, and shopping experience on purchase decision and its implications on customer satisfaction: Case study of Geraipedia Jabodetabek customers. *Primanomics: Jurnal Ekonomi & Bisnis*, 19(2), 46–61.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.
- Wedel, M., & Kamakura, W. A. (2012). *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). Springer.

Anexos

Anexo A. Diccionario de variables construidas

Tabla A.1: Variables principales construidas. Fuente: Elaboración propia.

Variable	Descripción
gasto_total	Suma del gasto registrado por hogar
gasto_per_capita	Gasto total dividido por número de personas
gasto_bebidas	Suma de gasto en productos clasificados como bebidas
share_bebidas	Gasto en bebidas dividido por gasto total
cantidad_total	Suma de cantidades registradas por hogar
n_productos	Número de glosas CCIF distintas compradas por hogar
n_establecimientos	Número de tipos de establecimiento distintos usados por hogar
share_supermercados	Gasto en supermercados dividido por gasto total
ingreso_per_capita	Ingreso disponible del hogar dividido por número de personas
edad_prom	Edad promedio de los integrantes del hogar
edue_prom	Escolaridad promedio de los integrantes del hogar
prop_menores_18	Proporción de integrantes menores de 18 años
prop_mayores_65	Proporción de integrantes de 65 años o más
clúster	Segmento asignado por K-Means

Anexo B. Código Python utilizado explicación por bloques

Este anexo presenta los bloques de código utilizados para generar los resultados presentados en la memoria, el objetivo de este código consiste en:

1. Cargar las bases base-personas-ix-epf.csv y base-cantidades-ix-epf.csv.
2. Limpiar, transformar e integrar ambas bases a nivel de hogar mediante folio_v.
3. Construir variables de gasto, ingreso, composición del hogar, productos, bebidas y canales de compra.
4. Generar análisis descriptivo, tablas y gráficos.
5. Entrenar modelos supervisados: Regresión Logística, KNN y SVM.
6. Evaluar modelos supervisados mediante matriz de confusión, Accuracy, Precision, Recall, Specificity, F1-score y ROC-AUC.
7. Aplicar K-Means como modelo no supervisado para segmentar hogares.
8. Evaluar K-Means mediante método del codo y Average Silhouette.
9. Guardar bases, tablas y gráficos.

Anexo B.1 Librerías

Este bloque importa las librerías necesarias para manipular datos, graficar, entrenar modelos de Machine Learning y calcular métricas de evaluación.

```
import os
import warnings
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder, label_binarize
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    confusion_matrix,
    classification_report,
    roc_auc_score,
    roc_curve,
    auc,
```

```

        silhouette_score
    )
    from sklearn.linear_model import LogisticRegression
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.svm import LinearSVC
    from sklearn.cluster import MiniBatchKMeans
    from sklearn.decomposition import PCA

```

Anexo B.2 Configuración de rutas

En este bloque se definen las rutas de entrada y salida. Para ejecutar el código en otro computador, dejar los archivos CSV en la misma carpeta que este script o modificar las rutas PERSONAS_PATH y CANTIDADES_PATH.

```

PERSONAS_PATH = "base-personas-ix-epf.csv"
CANTIDADES_PATH = "base-cantidades-ix-epf.csv"

```

```

# Carpeta donde se guardarán tablas, bases y gráficos generados.
OUT_DIR = "resultados_tesis"
FIG_DIR = os.path.join(OUT_DIR, "figuras")
os.makedirs(OUT_DIR, exist_ok=True)
os.makedirs(FIG_DIR, exist_ok=True)

```

Anexo B.3 Funciones auxiliares

Estas funciones permiten ordenar el código y reutilizar operaciones frecuentes, como limpiar números, formatear valores y crear gráficos.

```

def convertir_numerico(serie):
    """
    Convierte una columna a formato numérico.
    Es útil cuando los datos vienen con puntos como separador de miles o comas decimales.
    """
    return pd.to_numeric(
        serie.astype(str)
        .str.replace(".", "", regex=False)
        .str.replace(",", ".", regex=False),
        errors="coerce"
    )

def guardar_grafico(fig, nombre_archivo):
    """

```

Guarda un gráfico en formato PNG con resolución adecuada para insertar en Word o PowerPoint.

```
"""
ruta = os.path.join(FIG_DIR, nombre_archivo)
fig.tight_layout()
fig.savefig(ruta, dpi=180, bbox_inches="tight")
plt.close(fig)
return ruta
```

def grafico_barras_vertical(data, x, y, titulo, xlabel, ylabel, archivo, rotacion=0):

```
"""
Genera un gráfico de barras vertical para comparar valores promedio por grupo.
Se utiliza para gasto e ingreso promedio por macrozona y CSE.
"""
fig, ax = plt.subplots(figsize=(8, 4.8))
ax.bar(data[x].astype(str), data[y])
ax.set_title(titulo)
ax.set_xlabel(xlabel)
ax.set_ylabel(ylabel)
ax.tick_params(axis="x", rotation=rotacion)
ax.ticklabel_format(style="plain", axis="y")

# Etiquetas sobre cada barra para facilitar la lectura.
for i, valor in enumerate(data[y]):
    ax.text(i, valor, f'{valor:,.0f}'.replace(",", "."), ha="center", va="bottom", fontsize=8)

return guardar_grafico(fig, archivo)
```

def grafico_barras_horizontal_millones(data, categoria, valor, hogares, titulo, xlabel, ylabel, archivo):

```
"""
Genera gráfico de barras horizontal con eje X expresado en millones de CLP.
Además, agrega al final de cada barra el gasto total y la cantidad de hogares.
"""
plot_data = data.sort_values(valor).copy()
plot_data["valor_millones"] = plot_data[valor] / 1_000_000

fig, ax = plt.subplots(figsize=(9.5, 5.8))
ax.barh(plot_data[categoria].astype(str), plot_data["valor_millones"])
ax.set_title(titulo)
ax.set_xlabel(xlabel)
```

```

ax.set_ylabel(ylabel)
ax.grid(axis="x", alpha=0.25)

# Espacio adicional para etiquetas.
maximo = plot_data["valor_millones"].max()
ax.set_xlim(0, maximo * 1.22)

for _, row in plot_data.iterrows():
    etiqueta = f'{row["valor_millones"]:.1f} MM\_{int(row[hogares]):,}'
    hogares.replace(",", ".")
    ax.text(row["valor_millones"] + maximo * 0.015, row[categoria], etiqueta,
            va="center", fontsize=8)

return guardar_grafico(fig, archivo)

def calcular_specifcity_macro(matriz):
    """
    Calcula la especificidad promedio macro para una matriz de confusión multiclase.
    En multiclase se utiliza el enfoque One-vs-Rest: para cada clase, esa clase se
    considera positiva y todas las demás se consideran negativas.
    """
    especificidades = []
    for i in range(matriz.shape[0]):
        tp = matriz[i, i]
        fn = matriz[i, :].sum() - tp
        fp = matriz[:, i].sum() - tp
        tn = matriz.sum() - tp - fn - fp
        especificidad = tn / (tn + fp) if (tn + fp) > 0 else np.nan
        especificidades.append(especificidad)
    return np.nanmean(especificidades)

```

Anexo B.4 Carga de datos

Este bloque lee las dos bases CSV de la IX EP.

La base de personas contiene características sociodemográficas y económicas.

La base de cantidades contiene registros de gasto, cantidad, producto y establecimiento.

```

personas = pd.read_csv(PERSONAS_PATH, sep=";", thousands=".")
cantidades = pd.read_csv(CANTIDADES_PATH, sep=";", thousands=".")

# Conversión preventiva de variables relevantes a formato numérico.
for col in ["ing_disp_hog_hd", "ing_disp_hd", "edad", "edue", "npersonas", "cse",
            "macrozona"]:

```

```

if col in personas.columns:
    personas[col] = convertir_numerico(personas[col])

for col in ["gasto", "cantidad", "macrozona", "folio_v"]:
    if col in cantidades.columns:
        cantidades[col] = convertir_numerico(cantidades[col])

```

Anexo B.5 Construcción de variables de personas a nivel hogar

La unidad de análisis de la tesis es el hogar. Por eso, la base de personas se resume por folio_v. Se calculan edad promedio, escolaridad, composición etaria, proporción de mujeres, ingreso del hogar y datos del sustentador principal.

```

sustentador = personas[personas["sprincipal"] == 1].copy()
sustentador_hogar = sustentador.groupby("folio_v").agg(
    edad_sust=("edad", "first"),
    edue_sust=("edue", "first"),
    sexo_sust=("sexo", "first"),
    ing_disp_hd_sust=("ing_disp_hd", "first")
).reset_index()

personas_hogar = personas.groupby("folio_v").agg(
    macrozona=("macrozona", "first"),
    cse=("cse", "first"),
    npersonas=("npersonas", "first"),
    edad_prom=("edad", lambda x: x[x >= 0].mean()),
    edad_mediana=("edad", lambda x: x[x >= 0].median()),
    edad_max=("edad", lambda x: x[x >= 0].max()),
    edad_min=("edad", lambda x: x[x >= 0].min()),
    prop_mujeres=("sexo", lambda x: (x == 2).mean()),
    prop_menores_18=("edad", lambda x: (x.between(0, 17)).mean()),
    prop_mayores_65=("edad", lambda x: (x >= 65).mean()),
    edue_prom=("edue", lambda x: x[x >= 0].mean()),
    ing_disp_hog_hd=("ing_disp_hog_hd", "first"),
    ing_personas_suma=("ing_disp_hd", lambda x: x[x > 0].sum())
).reset_index()

personas_hogar = personas_hogar.merge(sustentador_hogar, on="folio_v", how="left")

```

Anexo B.6 Construcción de variables de consumo a nivel hogar

Este bloque transforma la base de cantidades a nivel hogar. Se calculan gasto total, cantidad total, número de productos, número de establecimientos, gasto en bebidas, gasto en alimentos y gasto por categorías amplias.

```

# Identificación de productos asociados a bebidas mediante palabras clave.
cantidades["is_bebida"] = cantidades["glosa_ccif"].astype(str).str.contains(

"BEBIDA|BEBIDAS|AGUA|JUGO|JUGOS|NÉCTAR|NECTAR|CAFÉ|CAFE|TÉ|TE|LEC
HE|CERVEZA|VINO|LICOR|ALCOHOL",
    case=False,
    regex=True,
    na=False
)

cantidades_hogar = cantidades.groupby("folio_v").agg(
    gasto_total=("gasto", "sum"),
    cantidad_total=("cantidad", "sum"),
    n_registros=("gasto", "size"),
    n_productos=("glosa_ccif", "nunique"),
    n_establecimientos=("glosa_establecimiento", "nunique"),
    gasto_prom_producto=("gasto", "mean"),
    gasto_mediano_producto=("gasto", "median"),
    gasto_max_producto=("gasto", "max"),
    cantidad_prom=("cantidad", "mean")
).reset_index()

bebidas_hogar = cantidades[cantidades["is_bebida"]].groupby("folio_v").agg(
    gasto_bebidas=("gasto", "sum"),
    cantidad_bebidas=("cantidad", "sum"),
    n_bebidas=("glosa_ccif", "nunique")
).reset_index()

alimentos_hogar = cantidades[~cantidades["is_bebida"]].groupby("folio_v").agg(
    gasto_alimentos=("gasto", "sum"),
    cantidad_alimentos=("cantidad", "sum"),
    n_alimentos=("glosa_ccif", "nunique")
).reset_index()

# Categorías amplias utilizadas para describir patrones de consumo.
patrones_categoria = {
    "gasto_pan_cereales": "PAN|ARROZ|CEREAL|PASTA|FIDEO|HARINA|GALLET",
    "gasto_lacteos": "LECHE|QUESO|YOGUR|YOGHURT|MANTEQUILLA",
    "gasto_carnes":
"CARNE|POLLO|VACUNO|CERDO|EMBUTIDO|JAM[ÓO]N|PESCADO|MARISCOS",

```

```

    "gasto_frutas_verduras":
    "FRUTA|VERDURA|PAPA|TOMATE|LIMON|LIMONES|PL[ÁA]TANO|MANZANA|L
    ECHUGA",
    "gasto_azucar_confites":
    "AZ[ÚU]CAR|CHOCOLATE|DULCE|HELADO|CONFITE|MERMELADA",
    "gasto_bebidas_azucaradas":          "BEBIDAS          GASEOSAS|BEBIDAS
    ENERGIZANTES|BEBIDAS ISOT[ÓO]NICAS|JUGOS|N[ÉE]CTAR",
}

```

```

categorias = []
for nombre_variable, patron in patrones_categoria.items():
    gasto_categoria = (
        cantidades[cantidades["glosa_ccif"].astype(str).str.contains(patron,
                                case=False,
                                regex=True, na=False)]
        .groupby("folio_v")["gasto"]
        .sum()
        .rename(nombre_variable)
    )
    categorias.append(gasto_categoria)

```

```

categorias_hogar = pd.concat(categorias, axis=1).fillna(0).reset_index()

```

Identificación del gasto realizado en supermercados.

```

supermercados_hogar = (
    cantidades[cantidades["glosa_establecimiento"].astype(str).str.contains("SUPERMERCAD
    O", case=False, na=False)]
    .groupby("folio_v")["gasto"]
    .sum()
    .rename("gasto_supermercados")
    .reset_index()
)

```

Anexo B.7 Integración final de bases y creación de ratios

Se unen las variables de personas y consumo mediante folio_v. Luego se crean variables per cápita y participaciones porcentuales.

```

df = (
    personas_hogar
    .merge(cantidades_hogar, on="folio_v", how="inner")
    .merge(bebidas_hogar, on="folio_v", how="left")
    .merge(alimentos_hogar, on="folio_v", how="left")
    .merge(categorias_hogar, on="folio_v", how="left")
)

```

```

        .merge(supermercados_hogar, on="folio_v", how="left")
    )

# Relleno de valores faltantes en variables de gasto/cantidad/conteo.
for columna in df.columns:
    if columna.startswith(("gasto_", "cantidad_", "n_")):
        df[columna] = df[columna].fillna(0)

# Ratios relevantes para análisis y modelamiento.
df["gasto_per_capita"] = df["gasto_total"] / df["npersonas"].replace(0, np.nan)
df["cantidad_per_capita"] = df["cantidad_total"] / df["npersonas"].replace(0, np.nan)
df["ingreso_per_capita"] = df["ing_disp_hog_hd"] / df["npersonas"].replace(0, np.nan)
df["share_bebidas"] = df["gasto_bebidas"] / df["gasto_total"].replace(0, np.nan)
df["share_supermercados"] = df["gasto_supermercados"] / df["gasto_total"].replace(0, np.nan)

for columna in patrones_categoria:
    nombre_share = "share_" + columna.replace("gasto_", "")
    df[nombre_share] = df[columna] / df["gasto_total"].replace(0, np.nan)

# Se conservan únicamente hogares con clasificación socioeconómica válida.
df = df[df["cse"].isin([1, 2, 3])].copy()
df.replace([np.inf, -np.inf], np.nan, inplace=True)

# Guardado de la base integrada final.
df.to_csv(os.path.join(OUT_DIR, "base_integrada_epf_hogares_ml.csv"), index=False)

```

Anexo B.8 Análisis descriptivo de datos

Este bloque genera estadísticos descriptivos generales, análisis por CSE, análisis por macrozona y matriz de correlación de variables principales.

```

variables_principales = [
    "gasto_total", "gasto_per_capita", "ing_disp_hog_hd", "ingreso_per_capita",
    "gasto_bebidas", "share_bebidas", "cantidad_total", "npersonas",
    "edad_prom", "edue_prom", "n_productos", "n_establecimientos"
]

descriptivos = df[variables_principales].describe(percentiles=[0.25, 0.50, 0.75, 0.90]).T.round(2)
descriptivos.to_csv(os.path.join(OUT_DIR, "tabla_5_1_estadisticos_descriptivos.csv"))

macro_map = {1: "Norte", 2: "Gran Santiago", 3: "Centro", 4: "Sur"}
cse_map = {1: "Bajo", 2: "Medio", 3: "Alto"}

```

```

df["macrozona_label"] = df["macrozona"].map(macro_map)
df["cse_label"] = df["cse"].map(cse_map)

macro_stats = (
    df.groupby(["macrozona", "macrozona_label"])
        .agg(
            hogares=("folio_v", "count"),
            gasto_promedio=("gasto_total", "mean"),
            ingreso_promedio=("ing_disp_hog_hd", "mean"),
            gasto_pc_promedio=("gasto_per_capita", "mean"),
            ingreso_pc_promedio=("ingreso_per_capita", "mean"),
            npersonas_prom=("npersonas", "mean")
        )
        .reset_index()
        .sort_values("macrozona")
    )
macro_stats.to_csv(os.path.join(OUT_DIR, "tabla_5_2_gasto_ingreso_macrozona.csv"),
index=False)

cse_stats = (
    df.groupby(["cse", "cse_label"])
        .agg(
            hogares=("folio_v", "count"),
            gasto_promedio=("gasto_total", "mean"),
            ingreso_promedio=("ing_disp_hog_hd", "mean"),
            gasto_pc_promedio=("gasto_per_capita", "mean"),
            ingreso_pc_promedio=("ingreso_per_capita", "mean"),
            npersonas_prom=("npersonas", "mean")
        )
        .reset_index()
        .sort_values("cse")
    )
cse_stats.to_csv(os.path.join(OUT_DIR, "tabla_5_3_gasto_ingreso_cse.csv"), index=False)

# Principales establecimientos por gasto total, incorporando registros y hogares únicos.
top_establecimientos = (
    cantidades.groupby("glosa_establecimiento")
        .agg(
            gasto_total=("gasto", "sum"),
            registros=("gasto", "size"),
            hogares=("folio_v", "nunique")
        )
        .reset_index()
    )

```

```

        .sort_values("gasto_total", ascending=False)
        .head(10)
    )
    top_establecimientos.to_csv(os.path.join(OUT_DIR,
"tabla_5_4_gasto_establecimientos_hogares.csv"), index=False)

# Principales categorías CCIF por gasto total, incorporando registros y hogares únicos.
top_categorias = (
    cantidades.groupby("glosa_ccif")
        .agg(
            gasto_total=("gasto", "sum"),
            registros=("gasto", "size"),
            hogares=("folio_v", "nunique"),
            cantidad_total=("cantidad", "sum")
        )
        .reset_index()
        .sort_values("gasto_total", ascending=False)
        .head(10)
    )
    top_categorias.to_csv(os.path.join(OUT_DIR, "tabla_5_5_categorias_gasto_hogares.csv"),
index=False)

# Matriz de correlación.
corr_vars = [
    "gasto_total", "gasto_per_capita", "gasto_bebidas", "share_bebidas",
    "cantidad_total", "npersonas", "edad_prom", "edue_prom", "ing_disp_hog_hd",
    "ingreso_per_capita", "n_productos", "share_supermercados",
    "prop_menores_18", "prop_mayores_65"
]
correlacion = df[corr_vars].corr(numeric_only=True)
correlacion.round(3).to_csv(os.path.join(OUT_DIR, "tabla_matriz_correlacion.csv"))

fig, ax = plt.subplots(figsize=(10, 8))
im = ax.imshow(correlacion, cmap="coolwarm", vmin=-1, vmax=1)
ax.set_xticks(range(len(corr_vars)), corr_vars, rotation=90, fontsize=8)
ax.set_yticks(range(len(corr_vars)), corr_vars, fontsize=8)
fig.colorbar(im, ax=ax, fraction=0.046, pad=0.04)
ax.set_title("Matriz de correlación de variables principales")
guardar_grafico(fig, "figura_5_7_matriz_correlacion.png")

```

Anexo B.9 Gráficos descriptivos del capítulo 5.1

Este bloque genera los gráficos de barras para el análisis descriptivo:

Gasto promedio por macrozona, gasto promedio por CSE, ingreso promedio por CSE, ingreso promedio por macrozona, gasto total por establecimiento y categorías por gasto total.

```
grafico_barras_vertical(  
    macro_stats, "macrozona_label", "gasto_promedio",  
    "Gasto promedio por macrozona", "Macrozona", "Gasto promedio hogar (CLP)",  
    "figura_5_1_gasto_promedio_macrozona.png"  
)  
  
grafico_barras_vertical(  
    macro_stats, "macrozona_label", "ingreso_promedio",  
    "Ingreso promedio por macrozona", "Macrozona", "Ingreso promedio hogar (CLP)",  
    "figura_5_2_ingreso_promedio_macrozona.png"  
)  
  
grafico_barras_vertical(  
    cse_stats, "cse_label", "gasto_promedio",  
    "Gasto promedio por clasificación socioeconómica", "CSE", "Gasto promedio hogar  
(CLP)",  
    "figura_5_3_gasto_promedio_cse.png"  
)  
  
grafico_barras_vertical(  
    cse_stats, "cse_label", "ingreso_promedio",  
    "Ingreso promedio por clasificación socioeconómica", "CSE", "Ingreso promedio hogar  
(CLP)",  
    "figura_5_4_ingreso_promedio_cse.png"  
)  
  
grafico_barras_horizontal_millones(  
    top_establecimientos,  
    categoria="glosa_establecimiento",  
    valor="gasto_total",  
    hogares="hogares",  
    titulo="Gasto total por establecimiento",  
    xlabel="Gasto total (millones de CLP)",  
    ylabel="Establecimiento",  
    archivo="figura_5_5_gasto_total_establecimiento.png"  
)  
  
grafico_barras_horizontal_millones(  

```

```

top_categorias,
categoria="glosa_ccif",
valor="gasto_total",
hogares="hogares",
titulo="Categorías por gasto total",
xlabel="Gasto total (millones de CLP)",
ylabel="Categoría CCIF",
archivo="figura_5_6_categorias_gasto_total.png"
)

```

Anexo B.10 Modelos Supervisados

Este bloque entrena tres modelos supervisados para predecir la clasificación socioeconómica del hogar: Regresión Logística, KNN y SVM. Variable objetivo: cse, con clases 1 = Bajo, 2 = Medio y 3 = Alto.

```

features_modelo = [
    "macrozona", "npersonas", "edad_prom", "prop_mujeres", "prop_menores_18",
    "prop_mayores_65", "edue_prom", "edad_sust", "edue_sust", "sexo_sust",
    "ing_disp_hog_hd", "ingreso_per_capita", "gasto_total", "gasto_per_capita",
    "cantidad_total", "cantidad_per_capita", "gasto_bebidas", "share_bebidas",
    "n_registros", "n_productos", "n_establecimientos", "share_supermercados"
] + list(patrones_categoria.keys()) + ["share_" + c.replace("gasto_", "") for c in
patrones_categoria]

```

```

X = df[features_modelo]
y = df["cse"].astype(int)
clases = [1, 2, 3]
etiquetas_clases = ["Bajo", "Medio", "Alto"]

```

```

num_features = [c for c in X.columns if c not in ["macrozona", "sexo_sust"]]
cat_features = ["macrozona", "sexo_sust"]

```

Compatibilidad entre distintas versiones de scikit-learn.

try:

```

    onehot = OneHotEncoder(handle_unknown="ignore", sparse_output=False)

```

except TypeError:

```

    onehot = OneHotEncoder(handle_unknown="ignore", sparse=False)

```

Pipeline de preprocesamiento:

- Variables numéricas: imputación por mediana y estandarización.

- Variables categóricas: imputación por moda y codificación One-Hot.

```

preprocesador = ColumnTransformer([
    ("num", Pipeline([

```

```

        ("imputer", SimpleImputer(strategy="median")),
        ("scaler", StandardScaler())
    ]), num_features),
    ("cat", Pipeline([
        ("imputer", SimpleImputer(strategy="most_frequent")),
        ("onehot", OneHotEncoder())
    ]), cat_features)
])

# Separación entrenamiento/prueba con muestreo estratificado para conservar la proporción
# de CSE.
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=42, stratify=y
)

modelos = {
    "Regresión Logística": LogisticRegression(max_iter=1000, class_weight="balanced",
        random_state=42),
    "KNN": KNeighborsClassifier(n_neighbors=11, weights="distance"),
    "SVM": LinearSVC(C=1.0, class_weight="balanced", random_state=42, dual=False,
        max_iter=3000)
}

resultados = []
matrices = {}
scores_roc = {}
predicciones = {}
pipelines_entrenados = {}

for nombre, modelo in modelos.items():
    pipe = Pipeline([
        ("pre", preprocesador),
        ("model", modelo)
    ])

    # Entrenamiento del modelo.
    pipe.fit(X_train, y_train)
    pipelines_entrenados[nombre] = pipe

    # Predicción sobre conjunto de prueba.
    y_pred = pipe.predict(X_test)
    predicciones[nombre] = y_pred

```

```

# Matriz de confusión.
matriz = confusion_matrix(y_test, y_pred, labels=clases)
matrices[nombre] = matriz

# ROC-AUC multiclase One-vs-Rest.
if hasattr(pipe.named_steps["model"], "predict_proba"):
    score = pipe.predict_proba(X_test)
    auc_macro = roc_auc_score(y_test, score, multi_class="ovr", average="macro")
elif hasattr(pipe.named_steps["model"], "decision_function"):
    score = pipe.decision_function(X_test)
    y_bin = label_binarize(y_test, classes=clases)
    auc_macro = np.mean([roc_auc_score(y_bin[:, i], score[:, i]) for i in range(len(clases))])
else:
    score = None
    auc_macro = np.nan

scores_roc[nombre] = score

# Indicadores de desempeño.
resultados.append({
    "Modelo": nombre,
    "Accuracy": accuracy_score(y_test, y_pred),
    "Precision_macro": precision_score(y_test, y_pred, average="macro",
zero_division=0),
    "Recall_macro": recall_score(y_test, y_pred, average="macro", zero_division=0),
    "Specificity_macro": calcular_specificity_macro(matriz),
    "F1_macro": f1_score(y_test, y_pred, average="macro", zero_division=0),
    "ROC_AUC_macro_OVR": auc_macro
})

# Guardado de métricas comparativas.
resultados_df = pd.DataFrame(resultados).round(4)
resultados_df.to_csv(os.path.join(OUT_DIR,
"tabla_5_6_metricas_modelos_supervisados.csv"), index=False)

# Reporte por clase: Precision, Recall, F1-score y soporte para Bajo, Medio y Alto.
reportes = []
for nombre, y_pred in predicciones.items():
    reporte = classification_report(
        y_test,
        y_pred,
        labels=clases,
        target_names=etiquetas_clases,

```

```

        output_dict=True,
        zero_division=0
    )
    for clase in etiquetas_clases:
        fila = reporte[clase].copy()
        fila["Modelo"] = nombre
        fila["Clase"] = clase
        reportes.append(fila)

pd.DataFrame(reportes).to_csv(os.path.join(OUT_DIR,
"reporte_clasificacion_por_clase.csv"), index=False)

```

Anexo B.11 Gráficos de matrices de confusión

Este bloque grafica la matriz de confusión de cada modelo supervisado. La matriz permite observar aciertos y errores por clase.

```

for nombre, matriz in matrices.items():
    fig, ax = plt.subplots(figsize=(6, 5))
    im = ax.imshow(matriz, cmap="Blues")
    ax.set_title(f"Matriz de confusión - {nombre}")
    ax.set_xlabel("Predicción")
    ax.set_ylabel("Real")
    ax.set_xticks(range(len(etiquetas_clases)), etiquetas_clases)
    ax.set_yticks(range(len(etiquetas_clases)), etiquetas_clases)
    fig.colorbar(im, ax=ax, fraction=0.046, pad=0.04)

    for i in range(matriz.shape[0]):
        for j in range(matriz.shape[1]):
            ax.text(j, i, matriz[i, j], ha="center", va="center", color="black", fontsize=11)

nombre_limpio = nombre.replace(" ", "_").replace("ó", "o").replace("í", "i")
guardar_grafico(fig, f"matriz_confusion_{nombre_limpio}.png")

```

Anexo B.12 Curva ROC del mejor modelo según ROC-AUC

Este bloque grafica la curva ROC One-vs-Rest para el modelo con mayor ROC-AUC. Esto permite analizar la capacidad discriminativa de cada clase.

```

mejor_modelo_roc = resultados_df.sort_values("ROC_AUC_macro_OVR",
ascending=False).iloc[0]["Modelo"]
score_mejor = scores_roc[mejor_modelo_roc]
y_test_bin = label_binarize(y_test, classes=clases)

```

```

if score_mejor is not None:
    fig, ax = plt.subplots(figsize=(7, 5))
    for i, clase in enumerate(etiquetas_clases):
        fpr, tpr, _ = roc_curve(y_test_bin[:, i], score_mejor[:, i])
        roc_auc = auc(fpr, tpr)
        ax.plot(fpr, tpr, label=f'{clase} AUC={roc_auc:.3f}')

    ax.plot([0, 1], [0, 1], linestyle="--", label="Azar")
    ax.set_title(f'Curva ROC One-vs-Rest - {mejor_modelo_roc}')
    ax.set_xlabel("Tasa de falsos positivos")
    ax.set_ylabel("Tasa de verdaderos positivos")
    ax.legend(loc="lower right")
    ax.grid(alpha=0.25)
    guardar_grafico(fig, "curva_roc_mejor_modelo.png")

```

Anexo B.13 Modelo No Supervisado K-means

Este bloque aplica K-Means para segmentar hogares sin usar la etiqueta CSE. Se utilizan variables de gasto, ingreso, bebidas, supermercados, composición del hogar, edad, escolaridad y categorías de alimentos.

```

variables_cluster = [
    "gasto_per_capita", "gasto_total", "gasto_bebidas", "share_bebidas",
    "n_productos", "share_supermercados", "ingreso_per_capita", "npersonas",
    "edad_prom", "edue_prom", "prop_menores_18", "prop_mayores_65"
] + list(patrones_categoria.keys())

X_cluster = df[variables_cluster].replace([np.inf, -np.inf], np.nan)
X_cluster = SimpleImputer(strategy="median").fit_transform(X_cluster)
X_cluster = StandardScaler().fit_transform(X_cluster)

```

Anexo B.14 Evaluación de K-means: Método del codo y Average silhouette

Se prueban valores de k entre 2 y 8.

Inercia: mide la suma de distancias cuadráticas al centroide. Menor inercia indica clusters más compactos.

Silhouette: evalúa cohesión y separación. Valores más cercanos a 1 indican mejor separación.

```

rng = np.random.default_rng(42)
idx_muestra = rng.choice(len(X_cluster), size=min(500, len(X_cluster)), replace=False)
X_sil = X_cluster[idx_muestra]

inercia = []

```

```

silueta = []

for k in range(2, 9):
    km = MiniBatchKMeans(n_clusters=k, random_state=42, batch_size=256, n_init=5,
max_iter=50)
    etiquetas = km.fit_predict(X_cluster)

    inercia.append({"k": k, "inertia": km.inertia_})
    silueta.append({"k": k, "silhouette": silhouette_score(X_sil, km.predict(X_sil))})

elbow_df = pd.DataFrame(inercia)
sil_df = pd.DataFrame(silueta)

elbow_df.to_csv(os.path.join(OUT_DIR, "tabla_5_7_kmeans_elbow.csv"), index=False)
sil_df.to_csv(os.path.join(OUT_DIR, "tabla_5_7_kmeans_silhouette.csv"), index=False)

# Gráfico del método del codo.
fig, ax = plt.subplots(figsize=(6.5, 4.5))
ax.plot(elbow_df["k"], elbow_df["inertia"], marker="o")
ax.set_title("Método del codo (Elbow) - K-Means")
ax.set_xlabel("Número de clusters (k)")
ax.set_ylabel("Inercia")
ax.grid(alpha=0.25)
guardar_grafico(fig, "figura_5_12_kmeans_elbow.png")

# Gráfico del Average Silhouette.
fig, ax = plt.subplots(figsize=(6.5, 4.5))
ax.plot(sil_df["k"], sil_df["silhouette"], marker="o")
ax.set_title("Método Average Silhouette - K-Means")
ax.set_xlabel("Número de clusters (k)")
ax.set_ylabel("Average Silhouette")
ax.grid(alpha=0.25)
guardar_grafico(fig, "figura_5_13_kmeans_silhouette.png")

```

Anexo B.15 Entrenamiento final de K-means con K=4

Según el análisis del método del codo, Average Silhouette e interpretabilidad de negocio, se utiliza una solución final de 4 clusters.

```

k_final = 4
kmeans_final = MiniBatchKMeans(n_clusters=k_final, random_state=42, batch_size=256,
n_init=5, max_iter=50)
df["cluster"] = kmeans_final.fit_predict(X_cluster) + 1

```

```
silhouette_final = silhouette_score(X_sil, kmeans_final.predict(X_sil))
```

```
perfil_clusters = df.groupby("cluster").agg(
    hogares=("folio_v", "count"),
    gasto_total_prom=("gasto_total", "mean"),
    gasto_pc_prom=("gasto_per_capita", "mean"),
    gasto_bebidas_prom=("gasto_bebidas", "mean"),
    share_bebidas_prom=("share_bebidas", "mean"),
    ingreso_pc_prom=("ingreso_per_capita", "mean"),
    npersonas_prom=("npersonas", "mean"),
    edad_promedio=("edad_prom", "mean"),
    edue_promedio=("edue_prom", "mean"),
    n_productos_prom=("n_productos", "mean"),
    share_super_prom=("share_supermercados", "mean"),
    cse_prom=("cse", "mean")
).round(2)
```

```
perfil_clusters.to_csv(os.path.join(OUT_DIR, "tabla_5_8_perfil_clusters.csv"))
df.to_csv(os.path.join(OUT_DIR, "base_integrada_epf_hogares_ml_clusters.csv"),
index=False)
```

Anexo B.16 Visualización PCA de clusters

PCA reduce las variables del modelo a dos componentes principales para poder visualizar gráficamente los clusters generados por K-Means.

```
pca = PCA(n_components=2, random_state=42)
componentes = pca.fit_transform(X_cluster)
```

```
fig, ax = plt.subplots(figsize=(6.5, 5.2))
scatter = ax.scatter(componentes[:, 0], componentes[:, 1], c=df["cluster"], s=8,
cmap="tab10", alpha=0.6)
ax.set_title("Visualización PCA de clusters K-Means")
ax.set_xlabel("Componente principal 1")
ax.set_ylabel("Componente principal 2")
fig.colorbar(scatter, ax=ax, label="Cluster")
guardar_grafico(fig, "figura_5_14_pca_clusters.png")
```