Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

Programación y recursos didácticos para la enseñanza de segmentación de mercados

Autora: Sandra Delgado Bratos
Tutor: Pablo Sánchez Mayoral
Año: 2025

A Bicho y a Tora,
porque me dieron aliento cuando más dudé de mí misma.

Agradecimientos

Quisiera dar las gracias a Pablo Sánchez Mayoral, mi tutor, por haberme guiado en este proyecto siempre con la mejor disposición. Agradezco a María Jesús y Pedro Luis, mis padres, y a Eduardo, mi hermano, que día a día han estado a mi lado deseándome lo mejor. Gracias también a mis compañeros, ahora amigos y más que amigos, con los que aprendí que las cosas se aprenden mejor cuando te las tomas a risa. Gracias a mis profesores de la carrera, que, además de cumplir con lo que está en las guías docentes, cada uno de ellos me enseñó a su manera qué esconden los números. No quiero olvidarme de agradecer a todos mis amigos de otros rubros, que piensan que la estadística es lo más complicado de este mundo y que siguen preguntándome e interesándose pese a ello. Y por último, pero no menos importante, gracias a Jesús Ángel, mi pareja, que ha sido mi mayor proveedor de chocolate y generador de momentos anti-estrés.

Resumen

En este proyecto se desarrolla un material didáctico integral para la enseñanza de la segmentación de mercados en la universidad, incorporando una plataforma educativa. Se organizó el contenido de manera progresiva para facilitar el aprendizaje, se diseñó un proyecto formativo completo con todos los recursos necesarios para llevar a cabo la labor de enseñanza, tanto teórica como práctica, y se estructuró la metodología de evaluación.

El enfoque utilizado para la creación del programa didáctico ha sido competencial, pensado para que los alumnos apliquen las técnicas estadísticas de segmentación de mercados a datos reales y para fomentar el pensamiento crítico.

Para el diseño de la plataforma educativa se ha usado un enfoque basado en microservicios, que facilita el mantenimiento y la evolución del sistema. Como sistema de contenerización se ha utilizado Docker y los principales servicios desplegados son Nextcloud, como sistema de almacenamiento, y JupyterLab, como entorno de trabajo.

Palabras clave: segmentación de mercados, material didáctico digital, plataforma educativa, técnicas estadísticas, desarrollo de recursos educativos.

Abstract

This project develops a comprehensive teaching resource for the instruction of market segmentation at the university level, incorporating an educational platform. The content was organized progressively to facilitate learning, and a complete instructional project was designed, including all the necessary resources to carry out both theoretical and practical teaching. The evaluation methodology was also structured.

The approach used in the creation of the teaching program is competency-based, aimed at enabling students to apply statistical market segmentation techniques to real data and to foster critical thinking.

For the design of the educational platform, a microservices-based approach was used, which facilitates system maintenance and evolution. Docker was used as the containerization system, and the main services deployed are Nextcloud, as a storage system, and JupyterLab, as a working environment.

Keywords: market segmentation, digital teaching material, educational platform, statistical techniques, educational resource development.

Índice

Capítulo 1.
Introducción

1.1. Antecedentes y motivación

En un entorno empresarial cada vez más competitivo y globalizado, las empresas deben comprender profundamente a sus consumidores para poder satisfacer sus necesidades de manera efectiva. La segmentación de mercados se erige como una herramienta indispensable en este proceso, permitiendo a las empresas dividir un mercado heterogéneo en grupos más pequeños y homogéneos de consumidores con características y necesidades similares. Esta práctica no solo facilita una mayor precisión en el diseño de estrategias de marketing, sino que también optimiza los recursos y maximiza la eficacia de las campañas publicitarias.

La segmentación de mercados comenzó a tomar forma en la década de 1950. Wendell R. Smith, en su artículo de 1956 "Product Differentiation and Market Segmentation as Alternative Marketing Strategies", (Smith, 1956), es ampliamente reconocido como uno de los pioneros que formalizó el concepto. Smith argumentó que las empresas podrían mejorar su efectividad al identificar y atender a grupos específicos de consumidores con características similares.

Durante las décadas de 1960 y 1970, la segmentación de mercados se consolidó como una práctica esencial en el marketing. Las empresas empezaron a utilizar variables demográficas (edad, género, ingresos), geográficas (región, clima), psicográficas (estilo de vida, personalidad) y conductuales (beneficios buscados, lealtad) para dividir sus mercados. Este periodo también vio el auge de las técnicas estadísticas y de investigación de mercados, que permitieron un análisis más detallado y preciso de los segmentos de consumidores.

Con el advenimiento de las tecnologías de la información y la comunicación en las décadas de 1980 y 1990, la segmentación de mercados se transformó aún más. La capacidad de recopilar y analizar grandes cantidades de datos llevó a enfoques más precisos y dinámicos, como la segmentación basada en datos de comportamiento en tiempo real y la segmentación personalizada. Las empresas comenzaron a utilizar bases de datos y software de análisis para crear perfiles detallados de sus clientes y ajustar sus estrategias de marketing en consecuencia.

En el siglo XXI, la segmentación de mercados ha evolucionado significativamente gracias a la digitalización y la explosión de los medios sociales y el comercio electrónico y las empresas pueden acceder a una gran cantidad de datos detallados sobre el comportamiento y las preferencias de los consumidores en línea. Esto ha dado lugar a la necesidad de crear sistemas de información para recopilar, almacenar, procesar y distribuir los datos de manera eficiente.

Consecuentemente, las empresas comenzaron a utilizar y especializar estos sistemas y llegaron la investigación de mercados y el marketing estratégico. La investigación de mercado implica el diseño, recogida y análisis de datos para resolver problemas específicos de marketing y el marketing estratégico se basa en segmentación de mercado, posicionamiento de producto y marketing mix para crear una ventaja competitiva sostenible.

Ambas prácticas son utilizadas me manera habitual por las empresas hoy en día y, por ello, la segmentación de mercados se ha consolidado como un componente crucial en los programas de estudios superiores en marketing y administración de empresas. Las universidades y escuelas de negocios incluyen este concepto en sus currículos, reconociendo su importancia para la formación de futuros profesionales del marketing.

Hoy en día, el punto de vista estadístico aporta rigor analítico, precisión y capacidad para manejar grandes conjuntos de datos, lo que resulta en una segmentación más eficaz y estratégicamente valiosa. Las habilidades en el desarrollo de modelos predictivos, validación de métodos y análisis multivariable son esenciales para una segmentación de mercados que realmente pueda transformar las estrategias de marketing y mejorar la competitividad de las empresas en un mercado global cada vez más complejo y dinámico.

1.2. Objetivos

El principal objetivo de este Trabajo de Fin de Grado (TFG) es desarrollar un documento exhaustivo que sirva como material didáctico integral y actualizado para la enseñanza de la segmentación de mercados en la universidad.

Este objetivo se desglosa en los siguientes puntos:

- Desarrollar los contenidos teóricos de forma estructurada y accesible.
- Establecer objetivos claros de aprendizaje para la materia.
- Proponer metodologías innovadoras que fomenten el aprendizaje significativo de los estudiantes.
- Diseñar un sistema de evaluación efectivo.
- Elaborar recursos prácticos que complementen y enriquezcan la comprensión de los conceptos teóricos.
- Incorporar herramientas digitales y tecnológicas que potencien la interactividad y la experiencia de aprendizaje.

1.3. Estructura de la memoria

Este documento se encuentra estructurado en cinco partes bien diferenciadas:

La primera, denominada FUNDAMENTOS TEÓRICOS, expone los conocimientos sobre la segmentación de mercados. Partiendo desde su concepto y utilidad en la investigación comercial y explicando las distintas metodologías a seguir para analizar la información y obtener la segmentación en base a técnicas estadísticas. Además, cuenta con ejemplos de aplicaciones reales.

La segunda parte, titulada El PROYECTO FORMATIVO DE LA MATERIA, concreta los objetivos de aprendizaje del alumno, los contenidos que se cubrirán, la metodología que seguiría el profesor y los recursos que se utilizarán para alcanzar dichos objetivos y una propuesta para evaluar los conocimientos adquiridos por los alumnos.

La tercera parte, nombrada RECURSOS TEÓRICOS PARA LA DOCENCIA, tiene como objetivo reunir todo aquello que un docente pudiera necesitar para la enseñanza de la segmentación de mercados en una clase teórica. Consta de las diapositivas en las que el docente puede apoyar sus explicaciones en el aula, así como de referencias.

La cuarta parte se denomina RECURSOS PRÁCTICOS PARA LA DOCENCIA. Al igual que en la anterior, reúne todo aquello necesario para un docente a la hora de enseñar de forma práctica los diferentes métodos de segmentación de mercados. Reúne tanto problemas resueltos como propuestos y la evaluación final.

En la quinta parte, llamada LA PLATAFORMA EDUCATIVA, se explica la principal herramienta digital pensada y creada como soporte a la enseñanza de los conocimientos expuestos en este trabajo. Contiene tanto su estructura como la guía de uso.

Finalizamos el documento con una CONCLUSIÓN, la cual engloba la consecución de objetivos propuestos, las principales dificultades enfrentadas, una evaluación del aprendizaje adquirido y las potenciales mejoras o ampliaciones del trabajo realizado.

Se incluye la BIBLIOGRAFÍA correspondiente y los ANEXOS pertinentes para complementar el contenido expuesto.

Capítulo 2.
Fundamentos teóricos

En esta sección se abordarán los fundamentos teóricos de la segmentación de mercados, proporcionando un marco conceptual que subraya su importancia en el ámbito del marketing. Se explorarán los principios y utilidades que sustentan la segmentación, así como los diferentes criterios utilizados para dividir un mercado en subgrupos homogéneos. Este análisis teórico es esencial para entender cómo y por qué aplicar la segmentación de mercados para mejorar la precisión de las estrategias de marketing, optimizar la satisfacción del cliente y lograr una ventaja en un entorno empresarial dinámico y competitivo.

En el contexto actual, caracterizado por la digitalización y la globalización, la información sigue siendo el bien más preciado para aquellos que quieran mantenerse presentes en el ámbito comercial. Se crearon, debido al avance de la tecnología y a la aparición de nuevas fuentes de información, la investigación de mercados y el concepto de marketing estratégico. Veremos brevemente estos términos:

2.1. Introducción a la segmentación de mercados

La investigación de mercados y el marketing estratégico son herramientas fundamentales para comprender las necesidades del consumidor. Su aplicación permite a las organizaciones adaptarse a un entorno dinámico y tomar decisiones más informadas.

Investigación de Mercado

La Investigación de Mercado (IM) es la obtención sistemática de información para asistir a la dirección en la toma de decisiones comerciales (Santesmases Mestre, 1996). Consiste en el diseño, recogida, análisis de datos e información relevante para resolver un problema concreto de marketing con el que se enfrenta la empresa, incluyendo información sobre los consumidores, competidores y el entorno general del mercado. Esta práctica permite a las empresas obtener datos precisos y relevantes que son fundamentales para la toma de decisiones informadas y estratégicas, lo que se conoce hoy en día como marketing estratégico.

La segmentación de mercados forma parte de este enfoque de marketing.

Marketing estratégico

El marketing estratégico es un enfoque a largo plazo en la planificación y ejecución de actividades de marketing que busca crear y mantener una ventaja competitiva sostenible. Se centra en identificar y satisfacer las necesidades y deseos del mercado objetivo de manera más efectiva que la competencia, alineando los recursos y capacidades de la empresa con las oportunidades del entorno del mercado. Involucra un análisis exhaustivo del entorno externo e interno, incluyendo factores económicos, sociales, tecnológicos, competitivos y regulatorios, para identificar oportunidades y amenazas.

Los pasos del marketing estratégico son:

1. **Análisis de la situación y Objetivos**

   Examinar el entorno interno y externo de la empresa para identificar fortalezas, debilidades, oportunidades y amenazas (análisis FODA). Además, se establecen los objetivos que la empresa desea alcanzar, los cuales deben ser específicos, medibles, alcanzables, relevantes y con un tiempo determinado (objetivos SMART).

2. **Segmentación de mercados**

   Se divide el mercado en grupos homogéneos de consumidores con características y necesidades similares. Veremos en más profundidad este punto a continuación.

3. **Posicionamiento de producto**

   Responde a cómo la empresa quiere que los consumidores perciban su producto o servicio en relación con la competencia. Se busca ocupar un lugar único y relevante en la mente de los consumidores, destacando atributos específicos que diferencien la oferta de la empresa.

4. **Marketing Mix**

   También conocido como las 4P del marketing (Producto, Precio, Plaza y Promoción). En este paso se toman las decisiones estratégicas sobre cómo la empresa va a ofrecer su producto o servicio al mercado. Se define el producto o servicio, su precio, los canales de distribución y las estrategias de promoción que se utilizarán para llegar a los consumidores.

5. **Implementación**

   Se traducen las estrategias en acciones concretas, asignando recursos, diseñando campañas publicitarias, capacitando al personal y poniendo en marcha todas las actividades necesarias para alcanzar los objetivos establecidos.

6. **Evaluación y Control**

   Por último, se monitorea y evalúa el desempeño de las estrategias de marketing implementadas. Se comparan los resultados obtenidos con los objetivos establecidos, se identifican desviaciones y se toman medidas correctivas si es necesario. El proceso de evaluación y control es continuo y permite a la empresa aprender de sus experiencias y mejorar constantemente sus acciones de marketing.

Con este contexto establecido, podemos profundizar en los fundamentos de la segmentación de mercados, un aspecto clave en la estrategia de marketing.

2.2. Fundamentos de la segmentación

La segmentación de mercados se fundamenta en la premisa de que los mercados no son homogéneos, sino que están compuestos por consumidores con diferentes necesidades, preferencias y comportamientos. Para abordar esta diversidad y optimizar sus estrategias de marketing, las empresas recurren a la segmentación de mercados, que se basa en seis principios clave:

1. **Diversidad del Consumidor**

   Los consumidores difieren en términos de características demográficas, geográficas, psicográficas y conductuales, que exploraremos más adelante. Reconocer y entender estas diferencias es fundamental para segmentar el mercado de manera efectiva.

2. **Optimización de Recursos**

   Al identificar y focalizarse en segmentos específicos, las empresas pueden asignar sus recursos de marketing de manera más eficiente. Esto significa que pueden diseñar productos, establecer precios, crear campañas de comunicación y elegir canales de distribución que se adapten específicamente a las características y necesidades de cada segmento, maximizando así el retorno de inversión.

3. **Mejora de la Satisfacción del Cliente**

   La segmentación permite a las empresas entender mejor las necesidades y deseos específicos de diferentes grupos de consumidores. Al personalizar sus ofertas y estrategias de marketing, las empresas pueden mejorar la satisfacción del cliente, fomentar la lealtad y, en última instancia, aumentar las tasas de retención.

4. **Diferenciación Competitiva**

   En un mercado competitivo, la capacidad de diferenciarse es crucial. La segmentación ayuda a las empresas a identificar nichos de mercado desatendidos o subvalorados y a desarrollar propuestas de valor únicas que atraigan a esos segmentos específicos, proporcionando una ventaja competitiva.

5. **Facilitación del Desarrollo de Productos**

   La segmentación proporciona información valiosa sobre las características y beneficios que los diferentes segmentos de mercado valoran. Esto facilita el desarrollo de productos y servicios que están alineados con las expectativas de cada segmento, aumentando las probabilidades de éxito en el mercado.

6. **Adaptabilidad y Respuesta al Cambio**

   Los mercados y los comportamientos de los consumidores cambian con el tiempo. La segmentación permite a las empresas ser más ágiles y adaptarse rápidamente a estos cambios al monitorear y analizar continuamente los diferentes segmentos del mercado, ajustando sus estrategias según sea necesario.

Definición de la segmentación de mercados

La segmentación de mercados es una herramienta del marketing estratégico que utiliza el análisis estadístico para dividir un mercado amplio y heterogéneo en subgrupos más pequeños y homogéneos. Su principal objetivo es maximizar la eficiencia y la efectividad de las actividades de marketing al adaptar productos, precios, comunicación y canales de distribución a las características únicas de cada segmento.

Ejemplos de la aplicación real

Veremos a continuación algunos ejemplos de cómo la segmentación de mercados fue utilizada por las empresas para adaptar sus productos, servicios y estrategias de marketing a las necesidades y preferencias específicas de diferentes grupos de consumidores, logrando así una mayor efectividad y éxito en sus mercados respectivos:

1. **Coca-Cola:** Adaptación geográfica y demográfica

   Segmenta según la ubicación y la demografía: lanza variantes de bebidas (Diet Coke, sabores regionales) y campañas específicas para públicos como jóvenes o mujeres. Por ejemplo, Coca-Cola Japón lanzó sabores como “Sakura” y “Matcha”, y en EEUU alinea campañas con eventos como el Super Bowl. (Zhang, 2024)

2. **Amazon**: Segmentación conductual (comportamiento de compra)

   Utiliza historial de navegación y compras previas para ofrecer recomendaciones personalizadas. Esto aumenta la retención y el valor de vida del cliente gracias a promociones y ofertas personalizados. (A & Syam, 2024)

3. **Intel**: Segmentación B2B con IA

   Desarrolló un sistema de segmentación automática que clasifica a empresas por industria y función para descubrir oportunidades de ventas, mejorando así la eficiencia de su fuerza comercial. (Lieder, Segal, Avidan, Cohen, & Hope, 2019)

4. **Salud pública en Malawi**: Segmentación según percepción de riesgo en VIH

   El proyecto BRIDGE en Malawi (2007) agrupó a casi 1 000 personas según su percepción del riesgo de contraer VIH y su creencia en poder prevenirlo. Resultaron cuatro segmentos: *responsive* (alto riesgo, alta eficacia), *avoidance* (alto riesgo, baja eficacia), *proactive* (bajo riesgo, alta eficacia) e *indifference* (bajo riesgo, baja eficacia). Así, se identificaron diferencias significativas en conocimientos sobre VIH, realización de test y uso de condón entre los grupos. (Rothman, Bartels, Wlaschin, & Salovey,, 2006)

5. **Hostelería**: Segmentación de clientes usando aprendizaje no supervisado

   Un estudio de 2022 propone una metodología con clustering jerárquico para segmentar a huéspedes según su comportamiento, características y tendencias temporales. Se creó un proceso escalable desde los datos crudos hasta su aplicación comercial, ayudando a las empresas hoteleras a ofrecer marketing personalizado y aumentar ingresos. (Kuo & Yang, 2022)

2.3. Principios para una segmentación de mercados

Después de haber explorado los fundamentos, la utilidad, la definición y algunos ejemplos ilustrativos de la segmentación de mercados, es momento de sumergirnos en las estrategias y metodologías específicas utilizadas para dividir un mercado heterogéneo en grupos más homogéneos y manejables.

Requisitos

Antes de llevar a cabo el proceso de segmentación de mercado, es crucial comprender los requisitos fundamentales que deben cumplirse para garantizar su efectividad y utilidad en el ámbito del marketing. En esta sección, exploraremos detalladamente los requisitos necesarios tanto para la segmentación en sí como para los segmentos de mercado resultantes, destacando la importancia de cada uno en el proceso de toma de decisiones estratégicas y la creación de valor para la empresa.

Requisitos básicos para practicar la segmentación

La segmentación de mercados requiere cumplir con una serie de criterios para garantizar su efectividad y utilidad en el ámbito del marketing. Estos requisitos fundamentales aseguran que la segmentación sea viable y conduzca a resultados positivos para la empresa:

- **Valoraciones Diferenciadas de los Consumidores**: Es crucial que existan variaciones en la percepción de los consumidores sobre aspectos como calidad, precio u otros atributos, dependiendo de la marca del producto. Estas diferencias proporcionan la base para la segmentación, ya que permiten identificar grupos de consumidores con necesidades y preferencias distintas.

- **Vinculación de las Diferencias Percibidas a Grupos de Consumidores**: Las diferencias percibidas deben ser lo suficientemente significativas como para agrupar a los consumidores en segmentos distintos. Sin esta vinculación no se podría conocer al público objetivo ni planificar una estrategia de marketing adecuada.
- **Posibilidad de Implementar Marketing Diferenciado**: Es esencial que sea factible diseñar y ejecutar estrategias de marketing distintas para cada segmento identificado, y que estas estrategias sean efectivas en la satisfacción de las necesidades y deseos específicos de cada grupo de consumidores.
- **Potencial de Ventas Suficiente**: Los segmentos deben tener un tamaño lo suficientemente grande y un potencial de ventas significativo como para justificar el esfuerzo y los recursos dedicados a su atención y satisfacción.
- **Estabilidad en el Tiempo**: Es deseable que los segmentos sean relativamente estables a lo largo del tiempo, lo que permite a las empresas desarrollar estrategias a largo plazo y obtener un retorno sostenible de sus inversiones en marketing.

Requisitos para elegir posibles segmentos objetivo

Para que un segmento de mercado sea efectivo y útil para la empresa, debe cumplir con una serie de requisitos específicos que faciliten su identificación, acceso y explotación adecuados:

- **Medible**: El segmento debe ser medible en términos de tamaño y potencial de compra. Es fundamental poder estimar el poder adquisitivo y el comportamiento de compra de los consumidores dentro del segmento.
- **Rentable**: El segmento debe ser lo suficientemente grande y homogéneo como para garantizar su rentabilidad. Debe representar una oportunidad significativa de ingresos para la empresa.
- **Accesible**: La empresa debe tener la capacidad de llegar efectivamente al segmento a través de sus canales de distribución y estrategias de marketing. De lo contrario, el segmento carecerá de valor práctico.
- **Factible**: La empresa debe ser capaz de gestionar y atender las necesidades del segmento de manera efectiva. Esto implica tener los recursos necesarios y la infraestructura adecuada para satisfacer las demandas del segmento de manera rentable.
- **Fácilmente Identificable**: El segmento debe poder ser fácilmente identificado y diferenciado de otros grupos de consumidores. Esto facilita la orientación de las estrategias de marketing y la personalización de las ofertas.
- **Heterogéneos**: Los consumidores dentro del segmento deben tener necesidades y características similares entre sí, pero diferentes de otros segmentos. Esto permite a la empresa diseñar productos y servicios que satisfagan de manera efectiva las necesidades específicas del segmento.
- **Defendible**: El segmento debe ser defendible frente a la competencia, lo que significa que la empresa debe ser capaz de mantener su posición en el mercado y proteger sus márgenes de beneficio dentro del segmento.

Criterios de segmentación

Los criterios de segmentación son las variables utilizadas para dividir el mercado en segmentos. Los podemos clasificar según si son generales o específicos, es decir, si son independientes del producto y/o del proceso de compra o si no. Además, podemos diferenciar si esta información es objetiva o subjetiva, dando lugar a la siguiente clasificación:

- **Criterios Generales Objetivos**
    - **Segmentación Demográfica y Socioeconómica**

      Basada en características como edad, género, ingresos, educación, ocupación y tamaño de la familia. Es uno de los métodos más comunes debido a la facilidad de obtención de estos datos y su relevancia para muchas decisiones de marketing.

    - **Segmentación Geográfica**

      Divide el mercado según regiones, ciudades, países, idioma, clima o densidad de población. Es útil para empresas que operan en múltiples áreas geográficas y necesitan adaptar sus estrategias a las peculiaridades locales.
- **Criterios Generales Subjetivos**
    - **Segmentación Psicográfica**

      Se centra en los aspectos psicológicos y de estilo de vida de los consumidores, como valores, intereses, actitudes y personalidad. Proporciona una visión más profunda de los motivadores de compra.

    - **Segmentación Conductual**

      Basada en el comportamiento de los consumidores en relación con el producto, incluyendo el uso del producto, la lealtad a la marca, la ocasión de uso y los beneficios buscados. Este enfoque es particularmente útil para identificar y atraer a los clientes más valiosos.
- **Criterios Específicos Objetivos**
    - **Segmentación por Uso**

      Este criterio se centra en cómo los consumidores utilizan el producto y puede incluir variables como la frecuencia de uso, el momento de uso y el lugar de uso. Es especialmente relevante para productos que se utilizan de manera diferente según el contexto o la situación.
- **Criterios Específicos Subjetivos**
    - **Segmentación por Beneficio**

      Se refiere a los beneficios percibidos por los consumidores al utilizar el producto o servicio. Esta segmentación se basa en los diferentes motivadores de compra y en cómo los consumidores valoran los distintos beneficios ofrecidos por el producto.

Variables

En la segmentación de mercados es fundamental comprender las variables que influyen en el comportamiento de compra de los consumidores.

Los dos tipos de variables más comunes según su naturaleza son:

- **Variables numéricas**

  También llamadas cuantitativas, son aquellas que expresan cantidades y se representan con números. Permiten realizar operaciones matemáticas como sumar o promediar.

  Ejemplos: edad, ingreso, número de hijos, temperatura.

  Se subdividen en:

    - **Discretas:** toman valores enteros (ej. número de productos comprados).

- **Continuas:** pueden tomar cualquier valor dentro de un rango (ej. estatura).

- **Variables categóricas**

  También llamadas cualitativas, son aquellas que representan cualidades o categorías, y no tienen un valor numérico intrínseco.

  Ejemplos: género, estado civil, tipo de producto.

  Se subdividen en:

  - **Nominales:** no tienen orden lógico (ej. color de ojos).
  - **Ordinales:** tienen un orden o jerarquía (ej. nivel de satisfacción: bajo, medio, alto).

Además, para llevar a cabo la segmentación de mercado se debe seleccionar adecuadamente la variable dependiente, que es nuestro objeto de estudio, e identificar las variables explicativas, que serán todas las demás. Podemos definirlas de la siguiente forma:

- **Variable Dependiente**

  La variable dependiente se refiere al aspecto específico que se investiga en el estudio. Seleccionar la variable indicada para el problema que se plantea es tan complejo como importante. Se debe partir de los objetivos del estudio y seleccionar cuidadosamente.

  Ejemplos: el consumo genérico, la preferencia por una marca en particular, el comportamiento de compra en diferentes contextos, la influencia del color en las decisiones de compra, o las características específicas de un producto o servicio que influyen en la decisión de compra.

- **Variables Explicativas**

  Las variables explicativas son los factores que se utilizan para explicar o predecir la variable dependiente. Pueden ser tantas variables como se considere necesario y cada una representa una información sobre los sujetos de estudio, sean clientes reales o potenciales o sean empresas. En general, estas variables son todos aquellos datos que corresponden a los criterios de segmentación.

Ambas, la variable dependiente y el conjunto de variables explicativas, pueden ser tanto variables numerales como variables categóricas. Dependiendo de su naturaleza será pertinente usar un método u otro.

Proceso de segmentación de mercados

El proceso de segmentación de mercados se compone de varios pasos clave que guían la identificación y comprensión de los diferentes grupos de consumidores y sus necesidades. Estos pasos son:

1. **Selección de Variable Dependiente**:

   En este primer paso, se determina la variable dependiente que servirá como base para la segmentación.

2. **Investigación y Análisis de Mercado**:

   Se procede a recopilar datos relevantes sobre los consumidores y el mercado en general. Esto implica la recopilación tanto de datos cuantitativos como cualitativos mediante diversas técnicas de investigación de mercado.

3. **Identificación de los Criterios de Segmentación**:

   Aquí se seleccionan las variables más apropiadas para segmentar el mercado, teniendo en cuenta los objetivos específicos de la empresa y las características del mercado objetivo. Este paso implica el uso de diversas técnicas de segmentación para determinar los criterios más relevantes.

4. **Desarrollo de Perfiles de Segmento**:

   Se procede a crear descripciones detalladas de cada segmento identificado con las técnicas de segmentación. Estos perfiles incluyen información sobre las características demográficas, psicográficas y comportamentales de los consumidores dentro de cada segmento.

5. **Evaluación y Selección de Segmentos**:

   En este paso, se analiza la viabilidad y atractivo de cada segmento identificado. Se consideran factores como el tamaño del segmento, su potencial de crecimiento y su nivel de competitividad en el mercado.

6. **Desarrollo de Estrategias de Marketing**:

   Finalmente, se formulan estrategias de marketing específicas para cada segmento seleccionado. Esto implica el diseño de productos y servicios adaptados a las necesidades de cada segmento, así como la implementación de estrategias de precios, comunicación y distribución adecuadas.

2.4. Técnicas estadísticas de segmentación

En esta sección, exploraremos las diversas técnicas de segmentación de mercado disponibles, proporcionando un amplio panorama de las herramientas que las empresas pueden utilizar para dividir el mercado en grupos homogéneos. Desde enfoques tradicionales hasta metodologías más avanzadas, analizaremos cada técnica, sobre su uso efectivo en la identificación y comprensión de los diferentes segmentos de mercado.

Comenzaremos con un esquema que ofrece una visión general de las diferentes técnicas que vamos a estudiar y a continuación detallaremos las técnicas.

2.4.1. Clasificación general

Para clasificar las técnicas de segmentación de mercados utilizaremos el criterio inicial de si el proceso es supervisado o no supervisado. Este criterio inicial distingue dos formas de trabajar con la información disponible.

I. Segmentación supervisada

Es un enfoque en el que se parte de información previa o etiquetas conocidas sobre los segmentos (como tipos de clientes o niveles de consumo). El objetivo es construir un modelo que pueda clasificar correctamente a nuevos individuos dentro de esos grupos establecidos.

II. Segmentación no supervisada

Es un enfoque que no utiliza etiquetas predefinidas. En lugar de ello, analiza los datos en busca de similitudes o patrones internos para formar grupos homogéneos de forma automática, sin conocimiento previo de cuántos o qué tipos de segmentos existen.

Quedando la clasificación de la siguiente forma:

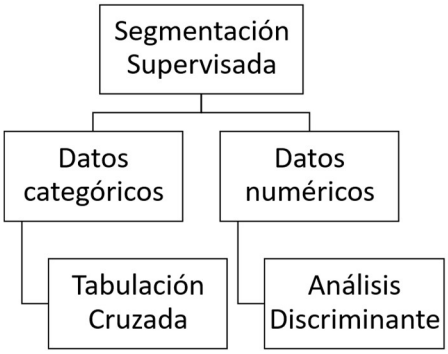


Ilustración 1. Clasificación de la segmentación supervisada

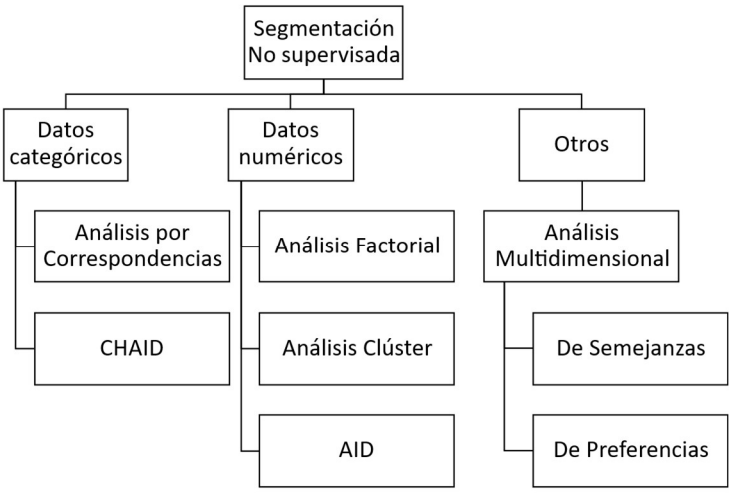


Ilustración 2. Clasificación de la segmentación no supervisada

2.4.2. Tabulación cruzada

La tabulación cruzada, también conocida como tabla de contingencia o tabla cruzada, es una técnica estadística utilizada para analizar la relación entre dos variables categóricas. Es la técnica más simple y fácil de aplicar, que además proporciona una clara visualización de la relación entre variables categóricas.

Cuando existen más de dos o tres criterios de clasificación, esta técnica empieza a no ser eficaz. Además, esta técnica no incorpora ningún tipo de contraste para estudiar estas diferencias, ni tampoco permite establecer jerarquías para determinar qué criterio tiene mayor poder discriminante sobre la población total, por lo que es preciso de realizarla junto con el test Chi-cuadrado.

El test Chi-cuadrado es una herramienta fundamental en la estadística inferencial, utilizada principalmente en el análisis de datos categóricos. Su propósito es evaluar si las diferencias observadas en una tabla de contingencia son debidas al azar o si existe una relación significativa entre las variables.

El test requiere un tamaño de muestra suficientemente grande (frecuencias esperadas en cada celda de 5 o más) y que las observaciones sean independientes.

Existen principalmente dos tipos de test chi cuadrado, pero el que nos interesa para la segmentación de mercados es el de independencia.

Aplicación:

Supongamos que queremos entender las preferencias de bebida de los clientes de una cafetería según su sexo, con el fin de diseñar promociones específicas. Para ello, realizamos una tabulación cruzada entre las variables 'sexo del comprador' y 'tipo de bebida comprada'.

Pasos a seguir:

1. **Obtención de los datos:** Por ejemplo, a partir de una encuesta.
2. **Tabla de frecuencias :** Las filas representan las categorías de una variable y las columnas representan las categorías de otra variable. Cada celda de la tabla muestra el conteo de casos que corresponden a la combinación particular de las categorías de las variables. Se añade la distribución de frecuencias de cada variable en la misma tabla. La tabla debe ser de n x m, siendo n el número de categorías de la variable de las filas y m el número de categorías de la variable de las columnas.
3. **Frecuencias porcentuales:** Siendo el 100% el número total de datos, se realiza una regla de tres para todos los valores de la tabla de frecuencias.
4. **Análisis de la tabla :** Anotar, para cada variable, qué categoría tiene mayor frecuencia y, en conjunto, qué combinación de categorías. Además, se identifican patrones y asociaciones significativas entre las variables.
5. **Representación:** Opcionalmente, se puede usar un gráfico de barras o gráfico de dispersión.
6. **Test Chi-cuadrado :**
    a. **Hipótesis**: Antes de realizar el test Chi cuadrado, es necesario establecer las hipótesis nula y alternativa. La hipótesis nula (H0) asume que no hay relación entre las variables, mientras que la hipótesis alternativa (H1) sugiere lo contrario. La no relación entre las variables indicaría independencia.
    b. **Cálculo de las frecuencias esperadas**:

    F esperada ij = Total i x Total j / Total

    c. **Cálculo del Estadístico Chi Cuadrado**: Utilizando la tabla de contingencia, calcula el estadístico chi cuadrado (χ^2) utilizando la fórmula:

    $\chi 2 = \sum[(Oij - Eij)2 / Eij]$

    Siendo:

    Eij = frecuencia esperada

    Oij = frecuencia observada

    d. **Cálculo de los Grados de Libertad:**

    (n_filas - 1) x(n_columnas - 1).

    e. **Determinación del valor crítico**: Calculado con una distribución chi cuadrado con los grados de libertad correspondientes.

    f. **Interpretación de los Resultados**: Basándose en el valor crítico y en el estadístico chi cuadrado, se interpreta si existe o no una asociación significativa entre las variables en estudio. Usaremos el nivel de significancia predefinido 0’05.

        - Si el estadístico es mayor que el valor crítico, se rechaza la hipótesis nula.
        - Si el estadístico es menor al valor crítico, decimos que no se puede rechazar la hipótesis nula, pues no hay diferencias significativas.

    g. **P-valor:** Calculado mediante software, es un valor entre 0 y 1 que indica la probabilidad de que la hipótesis nula es verdadera.

        - Si el p-valor es menor que 0.05, rechazamos la hipótesis nula.
        - Si el p-valor es mayor que 0.05, decimos que no se puede rechazar la hipótesis nula, pues no hay diferencias significativas.

7. **Resultados:** Con toda la información obtenida, podemos responder a si merece o no la pena segmentar la población en base a las dos variables seleccionadas. Si las variables son independientes, es decir, si no rechazamos la hipótesis nula del test Chi-cuadrado, segmentar la población no es interesante. En caso contrario, la segmentación de mercados sí es recomendable.

Continuando el ejemplo de aplicación:

La tabla obtenida nos permite identificar patrones. Por ejemplo, que los hombres prefieren el café, mientras que las mujeres consumen más té. Con base en esta información, podemos segmentar nuestras campañas: ofrecer descuentos en té durante horarios con mayor afluencia femenina o destacar el café en publicidad dirigida a clientes masculinos.

Esta técnica está indicada para analizar la relación entre dos variables categóricas, especialmente cuando se desea identificar si dos variables están relacionadas y cómo se cruzan las categorías. Es particularmente útil cuando los datos se presentan en forma de conteos o frecuencias, como número de compras, respuestas a encuestas, o selección de productos por grupos definidos (por ejemplo, edad, género, región o nivel socioeconómico).

2.4.3. Análisis Discriminante

El análisis discriminante es una técnica estadística utilizada para clasificar observaciones en diferentes grupos predefinidos, basándose en un conjunto de características predictoras. Es especialmente útil cuando queremos entender las diferencias entre varias categorías o clases y asignar nuevas observaciones a estas clases con base en sus características.

Existen varias variantes del análisis discriminante, siendo las más comunes el Análisis Discriminante Lineal (LDA, por sus siglas en inglés) y el Análisis Discriminante Cuadrático (QDA, por sus siglas en inglés).

El objetivo principal del LDA es encontrar una combinación lineal de variables predictoras que maximice la separación entre dos o más clases. Esto se logra buscando una proyección que maximice la distancia entre las medias de las clases y minimice la varianza dentro de cada clase.

Suposiciones:

1. **Normalidad Multivariada**: Las variables predictoras se distribuyen de manera normal dentro de cada clase.
2. **Igualdad de Covarianzas**: Las matrices de covarianza de las variables predictoras son iguales para todas las clases.
3. **Linealidad**: Las relaciones entre las variables predictoras y la variable de clase son lineales.

El QDA es una extensión del LDA que permite una mayor flexibilidad al relajar la suposición de igualdad de covarianzas entre las clases. Cada clase tiene su propia matriz de covarianza, lo que permite capturar relaciones no lineales entre las variables predictoras y la variable de clase. Tiene riesgo de sobreajuste, por lo que se recomienda usarla con más de

$$10 \times p\_variables \times k\_clases = nº \text{ de datos}$$

Aplicación:

Supongamos que queremos segmentar a nuestros clientes en distintos grupos de comportamiento de compra —por ejemplo, compradores frecuentes, cazadores de ofertas y clientes esporádicos— en función de variables observables como el número de compras mensuales, el porcentaje de descuento utilizado, y el gasto promedio por compra. El análisis discriminante nos permitiría construir una función que clasifique a nuevos clientes en uno de esos segmentos basándose en sus características.

Pasos a seguir:

1. **Definición de Grupos o Categorías**: Se identifican y definen las categorías o grupos en los que se desea clasificar los casos.
2. **Selección de Variables Predictoras**: Se seleccionan las variables predictoras que se utilizarán para distinguir entre los grupos.
3. **División de los Datos**: Se divide el conjunto de datos en dos: un conjunto de datos de entrenamiento y un conjunto de datos de prueba. Ambos conjuntos contienen la información de pertenencia a las categorías.
4. **Entrenamiento del Modelo**: Se utilizan los datos de entrenamiento para entrenar el modelo de análisis discriminante. Durante este proceso, el modelo busca la combinación óptima de variables predictoras que maximiza la separación entre los grupos definidos.

   Dado un conjunto de datos con n observaciones y p variables predictoras, el análisis discriminante lineal busca un vector de pesos w tal que la proyección w^T*x.

   - **LDA:**
     La función discriminante para una clase k se define como:
     $\delta k(x)=x^T*\Sigma-1\mu k-21\mu k^T\neg*\Sigma-1\mu k+\log(\pi k)$
     donde:
     - mu_k es el vector de medias de la clase k,
     - Sigma es la matriz de covarianza común,
     - pi_k es la probabilidad a priori de la clase k.

     Una nueva observación x se asigna a la clase k que maximice delta_k(x).
   - **QDA:**
     La función discriminante para el QDA se define como:
     $\delta k(x)=-21\log|\Sigma k|-21(x-\mu k)T\Sigma k-1(x-\mu k)+\log(\pi k)$
     donde:
     - Sigma_k es la matriz de covarianza de la clase k.

     Una nueva observación x se asigna a la clase k que maximice delta_k(x).
5. **Validación del Modelo**: Se valida la eficacia del modelo utilizando los datos de prueba. Esto ayuda a garantizar que el modelo pueda generalizarse a nuevos datos y no esté sobreajustando a la muestra de entrenamiento.
6. **Interpretación de los Resultados**
   - Coeficientes del modelo: identifican la capacidad de discriminación de cada variable y su relación con la clase.
   - Score de la función discriminante: comprobar que las clases estén bien separadas.
   - Matriz de confusión: determina la capacidad de clasificación del modelo.
   - Probabilidad a posteriori: seguridad de cada predicción.

7. **Clasificación de Nuevos Casos**: El modelo entrenado se puede utilizar para clasificar nuevos casos en las categorías definidas previamente.

Esta técnica está indicada para situaciones en las que el objetivo es predecir una variable categórica (grupos a priori) a partir de variables numéricas. Es especialmente útil cuando queremos entender qué variables permiten diferenciarlos mejor, y cómo se pueden asignar nuevos casos a esos grupos.

2.4.4. Análisis Factorial

El Análisis Factorial es una técnica estadística utilizada para identificar estructuras subyacentes en un conjunto de datos. Su objetivo principal es reducir la dimensionalidad. Este método es particularmente útil cuando se trabaja con datos de encuestas, cuestionarios y otros tipos de datos complejos donde hay múltiples variables que podrían estar interrelacionadas.

Aplicación:

Supongamos que queremos segmentar un grupo de consumidores en función de sus comportamientos de compra, como la frecuencia con que adquieren productos, su nivel de gasto mensual, la sensibilidad al precio, el uso de descuentos y el interés por productos premium. Con tantas variables involucradas, puede resultar complejo identificar patrones claros. El análisis factorial nos permite reducir esta complejidad al identificar factores subyacentes que explican la correlación entre esas variables, facilitando así la creación de segmentos de consumidores con características similares.

Pasos a seguir:

1. **Obtención de los datos**: Estos datos pueden ser obtenidos a través de encuestas, registros de ventas, análisis de redes sociales, u otras fuentes de información.
2. **Selección de variables**: Selecciona las variables que se utilizarán en el análisis.
3. **Evaluación de la adecuación de los datos**: Antes de realizar el análisis factorial, es importante evaluar la adecuación de los datos para este propósito. Las dos pruebas principales para esto son:
    - Prueba de Kaiser-Meyer-Olkin (KMO):
        - KMO > 0.8, excelente.
        - KMO entre 0.6 y 0.8, aceptable.
        - KMO < 0.5, inadecuado.
    - Prueba de esfericidad de Bartlett:
        - Si $p < 0.05$, los datos son adecuados.
4. **Cálculo de la Matriz de Correlación:** Calcular la matriz de correlación entre todas las variables para identificar posibles relaciones**.**
5. **Extracción de Factores**: Determinar el número de factores a extraer utilizando criterios como el criterio de Kaiser (factores con eigenvalores mayores a 1) o el análisis del scree plot.
6. **Rotación de Factores**: Aplicar una rotación para simplificar y clarificar la interpretación en caso de que se extraigan 2 o más factores.

- Varimax : la más usada cuando los factores no están correlacionados o se duda de ello.
- Oblimin : si se traba de constructos psicológicos, donde es normal que los factores se influyan entre sí.

7. **Interpretación de los resultados**: Identificar los factores extraídos y su carga factorial para las variables.
8. **Resultados**: Finalmente, se decide la segmentación a aplicar.

Esta técnica está indicada para situaciones en las que se dispone de múltiples variables relacionadas entre sí, y se desea resumir la información en un conjunto más pequeño de dimensiones interpretables (factores), que capturan la esencia del comportamiento observado. Los datos pueden ser tanto categóricos como numéricos.

2.4.5. Análisis por Correspondencias

El Análisis de Correspondencias (CA) es una técnica estadística multivariada diseñada para explorar y visualizar las relaciones entre categorías de variables cualitativas. A partir de una tabla de contingencia, CA transforma los datos en una representación gráfica en un espacio de dimensiones reducidas. En este espacio, cada categoría es representada como un punto, y la proximidad entre puntos refleja la intensidad de su asociación. Esto permite identificar segmentos de mercado con comportamientos similares, detectar asociaciones entre productos y perfiles de clientes.

El CA es una ampliación de la tabulación cruzada con técnicas del análisis factorial.

Los objetivos del Análisis de Correspondencias son:

- Reducir la dimensionalidad
- Representar gráficamente las relaciones entre categorías de las variables
- Identificar asociaciones y patrones.

Aplicación:

Supongamos que queremos entender las preferencias de distintos perfiles de consumidores en relación con los tipos de productos que compran en distintas regiones del país. El objetivo es descubrir patrones de asociación entre segmentos de clientes y tipos de productos para diseñar estrategias de marketing específicas por grupo.

Pasos a seguir:

1. **Creación de la tabla de contingencia:** Una tabla de contingencia consiste en representar datos categóricos en términos de conteos de frecuencia. Se construye una tabla de contingencia N, donde las filas representan las categorías de una variable X(1..i) y las columnas representan las categorías de la otra variable Y(1..j). Cada celda n_ij contiene la frecuencia de la combinación de categorías i y j.

2. **Aplicar el Análisis de Correspondencias:** Se aplica el algoritmo a la tabla N.

3. **Cálculo de las puntuaciones de correspondencia:** Se calculan las puntuaciones de correspondencia, que son medidas de asociación entre las categorías de las diferentes variables. Estas puntuaciones indican la fuerza y dirección de la relación entre las categorías.
4. **Construcción del Mapa de Correspondencias:** Se proyectan las categorías de filas y columnas en un espacio bidimensional o tridimensional. Este gráfico, conocido como mapa de correspondencias, visualiza las relaciones entre las categorías.
5. **Interpretación del Mapa de Correspondencias**
    a. Proximidad: Las categorías que están más cerca en el gráfico están más asociadas. Por ejemplo, si una categoría de la variable de fila está cerca de una categoría de la variable de columna, esto indica una asociación fuerte entre esas dos categorías.
    b. Distancias: La distancia entre puntos en el gráfico refleja la disimilitud entre las categorías. Cuanto más separados estén dos puntos, menos asociadas están las categorías correspondientes.
    c. Contribuciones: Se pueden calcular las contribuciones de cada categoría a las dimensiones principales, lo que permite entender qué categorías son más influyentes en la estructura de datos.

Esta técnica está indicada para analizar relaciones entre variables categóricas en forma de tablas de contingencia. A través de un mapa de correspondencias, el análisis permite representar visualmente las asociaciones más fuertes entre categorías, identificar grupos con comportamientos similares y facilitar la segmentación de mercado basada en afinidades observadas.

Es posible analizar más de dos variables categóricas a la vez con esta técnica. Sería necesario utilizar técnicas más avanzadas según el tipo de tabla de contingencia: Análisis de Correspondencias Múltiples en el caso de tener una única tabla de contingencia, o Análisis de Correspondencia Múltiple por Bloques si lo que se tiene son varias tablas de contingencia representando distintas combinaciones de variables.

2.4.6. Clúster Analysis

El análisis clúster, también conocido como análisis de conglomerados, es una técnica de aprendizaje no supervisado utilizada para agrupar un conjunto de objetos de tal manera que los objetos en el mismo grupo (o clúster) sean más similares entre sí que aquellos en otros grupos. La similitud se mide generalmente utilizando métricas de distancia, como la distancia euclidiana, aunque existen muchas otras opciones dependiendo del contexto y la naturaleza de los datos. Esta técnica es fundamental para comprender la estructura subyacente de conjuntos de datos complejos e identificar patrones.

Algunos Tipos de Clustering

1. Clustering Particional: Divide el conjunto de datos en un número predefinido de clústeres. El algoritmo K-means es el más conocido en esta categoría.
2. Clustering Jerárquico: Crea una jerarquía de clústeres que pueden representarse en un dendrograma. Puede ser aglomerativo (comienza con cada objeto como un clúster individual y fusiona los más cercanos) o divisivo (comienza con un único clúster y divide iterativamente). El más común es el llamado método de enlace (Linkage Method), es aglomerativo y, según la medición que haga, va uniendo los clústeres.

3. Modelos Basados en Densidad: Encuentran clústeres basados en áreas de alta densidad de puntos, como DBSCAN.
4. Modelos Basados en Distribución: Supone que los datos se generan a partir de una mezcla de distribuciones probabilísticas y agrupa los datos en consecuencia, como en el caso de Gaussian Mixture Models (GMM).

Algoritmos de Clustering que vamos a ver:

1. K-means:
   - Divide los datos en k clústeres.
   - Minimiza la suma de las distancias cuadradas entre los puntos y el centroide del clúster.
   - Es rápido y fácil de implementar, pero puede converger a un óptimo local.

2. Ward:
   - Es un método de enlace.
   - No requiere el número de clústeres de antemano.
   - Produce un dendrograma que muestra la fusión o división de clústeres.

Aplicación:

Supongamos que queremos segmentar a los clientes de una tienda de moda en línea con el objetivo de diseñar campañas personalizadas. Podemos aplicar el método de *K-Means* si ya tenemos una idea aproximada del número de segmentos que queremos analizar. Sin embargo, si no sabemos cuántos grupos existen naturalmente en nuestros datos, podemos utilizar el *clustering jerárquico* con el método de Ward para explorar visualmente cómo se forman los clústeres mediante un dendrograma y decidir a partir de allí cuántos segmentos tienen sentido comercialmente.

Pasos a seguir:

1. **Selección de variables**: Identificar las variables relevantes para el análisis.
2. **Preprocesamiento de datos**: En este paso se estandarizan las variables para asegurar que todas tengan el mismo peso en el análisis y se depuran los datos.
3. **Elección del método de clustering**: La elección del método y configuración de los parámetros dependiendo de la naturaleza de los datos y los objetivos del análisis.
   - **Definiendo un número de clústeres:** K-means
   - **Sin definir el número de clústeres:** Ward
4. **Validar los clústeres:** comprobar que el número es el adecuado y que están suficientemente diferenciados.
   - Método del codo: gráfico. Se calcula el ECM dentro del clúster para diferentes valores de k.
   - Coeficiente silueta: Compara la cohesión entre dentro del clúster con la separación del resto de datos para diferentes k. El valor más cercano a 1 es el mejor.
5. **Implementación del análisis**: Aplicar el algoritmo de clustering seleccionado.
6. **Gráficos y su interpretación**:

- **K-means**: Gráfico de dispersión
- **Ward:** Dendrograma

7. **Interpretación de resultados**: Analizar los clústers identificados para comprender las características distintivas de cada grupo.

K-Means es especialmente útil cuando se necesita segmentar grandes volúmenes de datos rápidamente y con alta eficiencia, siempre que las variables sean numéricas y estén debidamente normalizadas.

El método de Ward es más adecuado cuando se busca comprender la estructura jerárquica de los datos y se dispone de una base más pequeña o moderada, ya que permite visualizar cómo se agrupan los datos progresivamente y sin necesidad de definir el número de segmentos de antemano.

2.4.7. Análisis Multidimensional

El análisis multidimensional (MDS, por sus siglas en inglés) es una técnica estadística utilizada para representar visualmente objetos en un espacio de pocas dimensiones, preservando lo mejor posible sus relaciones de semejanza o disimilitud. Esta metodología se basa en la idea de que los individuos perciben las características de manera subjetiva y que estas percepciones pueden ser representadas en un espacio multidimensional.

Existen dos tipos principales de análisis:

- MDS métrico: las disimilitudes tienen significado numérico y trata de conservar las distancias reales entre objetos.
- MDS no métrico: se basa en el orden o ranking de las disimilitudes, sin suponer proporciones exactas.

Tipos de datos de entrada:

- MDS de Semejanza: los datos son pares de productos, marcas o atributos.
- MDS de Preferencia: los datos son preferencias individuales respecto a un conjunto de productos, marcas o atributos.

Aplicación:

Supongamos que queremos saber la opinión de los consumidores y utilizaremos datos de semejanza o preferencia. Para ello hacemos una encuesta en la que los consumidores indiquen: cuán parecidos o diferentes son los productos/marcas entre sí (Semejanza), o qué producto/marca prefieren (Preferencia).

Pasos a seguir:

1. **Matriz de disimilitudes**: Se comienza con una matriz de disimilitudes entre los objetos. Los datos pueden provenir de encuestas, experimentos o cualquier otra fuente donde se evalúen comparativamente los objetos. Debe de ser una matriz cuadrada con ceros en su diagonal.
    a. **MDS no métrico**: en caso de que los datos iniciales sean ordinales, se transforman usando la distancia euclidiana.
2. **Aplicación del MDS**: aplicamos el análisis multidimensional. Usamos 2 dimensiones para una mejor visualización.

3. **Visualización**: Los resultados se visualizan en un diagrama de dispersión donde los objetos similares aparecen cerca unos de otros y los diferentes están más alejados. Esto facilita la interpretación visual de las relaciones complejas entre los objetos.

Esta técnica está indicada para datos de entrada subjetivos o de tipo ordinal. El MDS no métrico no asume que la diferencia entre categorías sea uniforme, solo que exista un orden en la percepción. Esto permite construir mapas perceptuales que reflejan el espacio mental del consumidor y descubrir segmentos basados en grupos de marcas similares.

2.4.8. CHAID

La técnica AID (Automatic Interaction Detection) es un algoritmo de árbol de decisión que parte de una variable dependiente numérica y utiliza análisis de varianza (ANOVA) para dividir los datos en grupos que expliquen mejor la variabilidad de dicha variable, generando divisiones binarias. AID busca maximizar la diferencia en promedios numéricos. Un ejemplo de uso es la predicción del gasto que haría un cliente.

CHAID (Chi-squared Automatic Interaction Detection) se deriva de AID como una especificación adaptada a variables dependientes categóricas, sustituyendo el ANOVA por pruebas de chi-cuadrado y permitiendo divisiones múltiples por nodo. CHAID busca identificar segmentos significativamente distintos en su distribución de categorías, siendo especialmente útil en segmentación de mercados y análisis de comportamiento.

Aplicación:

Supongamos que queremos segmentar a los clientes de una tienda en línea para identificar grupos con diferentes probabilidades de compra según características como edad, género, nivel de ingresos y hábitos de navegación. Aplicando CHAID, podemos dividir a los clientes en segmentos basados en combinaciones significativas de estas variables categorizadas, revelando, por ejemplo, que “mujeres jóvenes con ingresos medios que visitan la sección de tecnología tienen una alta probabilidad de comprar”.

Pasos a seguir:

1. **Recopilación y preparación de datos:** Se recopilan y preparan los datos relevantes para el análisis. Esto implica seleccionar las variables predictoras y la variable de interés, así como limpiar los datos para eliminar valores atípicos o faltantes que puedan afectar la precisión del modelo.
2. **Construcción del árbol:** CHAID utiliza un enfoque de árbol de decisión para segmentar los datos. Comienza dividiendo la muestra total en grupos homogéneos basados en la variable independiente que mejor predice la variable dependiente. Continúa dividiendo cada grupo en subgrupos en función de otras variables independientes, utilizando pruebas estadísticas como la prueba de chi-cuadrado para determinar la significancia de cada división.
3. **Comprobación:** Para comprobar la robustez del árbol creado, se comprueban el p-valor de cada división. Si es menor a 0.01, podemos decir que es robusta.
4. **Podado del árbol:** Una vez que se completa la construcción del árbol, se puede realizar el podado para evitar el sobreajuste del modelo. Esto implica eliminar nodos o ramas del árbol que no contribuyen significativamente a la precisión del modelo.

5. **Interpretación de resultados:** Finalmente, se interpreta el árbol de decisión resultante para comprender cómo las variables predictoras interactúan entre sí y cómo influyen en la variable de interés.

Esta técnica está indicada para cuando se busca una segmentación clara y fácilmente interpretable, especialmente en problemas donde la variable objetivo es categórica (como compra sí/no, preferencia de producto, o nivel de satisfacción). CHAID es útil para descubrir relaciones y grupos significativos dentro de grandes bases de datos, ayudando a diseñar estrategias de marketing personalizadas y basadas en evidencia estadística.

2.4.9. Otros modelos

En el ámbito de la segmentación de mercados, hay muchos otros modelos y enfoques que no se han mencionado anteriormente. Algunos de estos modelos incluyen:

1. **Análisis de Componentes Principales, Análisis Canónico, DBSCAN, t-SNE, UMAP:** son técnicas más avanzadas de segmentación no supervisada.
2. **CART, Probit, Logit, Random Forest, XGBoost, SVM, Modelos predictivos**: son técnicas más avanzadas de segmentación supervisada.
3. **Latent Class Analysis (LCA), Gaussian Mixture Models (GMM), Naive Bayes, Hidden Markov Models:** son modelos probabilísticos y latentes. Estas técnicas asumen que los datos provienen de distribuciones subyacentes no observables (latentes). En vez de asignar un caso a un único segmento, asignan probabilidades de pertenencia a varios grupos, lo que las hace ideales cuando los segmentos no son rígidos.
4. **RFM (Recency, Frequency, Monetary), Customer Lifetime Value (CLV), Propensity Modeling, Cohort Analysis:** son técnicas de segmentación conductual.
5. **Self-Organizing Maps (SOM), Autoencoders + Clustering, Deep Embedding Clustering:** técnica de deep learning y redes neuronales. Son muy potentes para datos no estructurados (imágenes, texto) o cuando las relaciones no son lineales.
6. **Análisis de sentimiento, Topic Modeling (LDA), BERT + clustering, Perceptual Mapping basado en texto:** es segmentación basada en texto y percepciones (NLP)**.** Estas técnicas permiten segmentar en función de opiniones, percepciones y lenguaje, ya sea en encuestas, redes sociales o reseñas. Son ideales para analizar la voz del cliente y extraer insights no estructurados.
7. **Técnicas mixtas** como PCA + K-means, RFM + Clustering, Autoencoders + DBSCAN, LDA + Segmentación por interés**.**

Capítulo 3.
El proyecto formativo de la materia

Este capítulo expone el diseño de lo que sería la guía docente para una asignatura, o parte de ella, dedicada a Segmentación de Mercados, desarrollada en el marco de este TFG. La propuesta presenta un enfoque teórico-práctico que integra técnicas estadísticas y herramientas digitales para identificar segmentos de mercado estratégicos. La estructura está orientada a fomentar el aprendizaje activo y la aplicación de conocimientos mediante ejercicios prácticos, en coherencia con los objetivos del programa y las demandas del entorno profesional actual.

3.1. Objetivos formativos

Conceptuales:

- Entender el concepto de segmentación de mercados y su relevancia estratégica en marketing.
- Conocer las diferentes técnicas de segmentación de mercados, así como sus fortalezas y limitaciones según el tipo de datos y el contexto del mercado.

Prácticos:

- Adquirir habilidades para aplicar técnicas de segmentación de mercado a través del uso de software especializado.
- Interpretar los resultados de análisis de segmentación y tomar decisiones basadas en ellos para seleccionar el segmento objetivo más adecuado.
- Desarrollar capacidades para comunicar de manera efectiva los hallazgos y conclusiones de los análisis de segmentación.

Transversales:

- Capacidad de trabajar en equipo y resolver problemas en un entorno de análisis de datos.
- Habilidades de comunicación, tanto escrita como oral, aplicadas a la presentación de resultados de análisis de mercado.

3.2. Contenidos de la materia

Los temas a desarrollar son:

1. **Introducción a la Segmentación de Mercados**

   Investigación de mercado. Marketing estratégico.

2. **Fundamentos de la Segmentación**

   Definición de la segmentación de mercados. Ejemplos de la aplicación real.

3. **Principios para una Segmentación de Mercados**

   Requisitos básicos para practicar la segmentación. Requisitos para elegir posibles segmentos objetivos. Criterios de segmentación de mercado. Variables para la segmentación de mercados. Proceso de segmentación de mercados.

4. **Técnicas Estadísticas de Segmentación**

   Clasificación general. Tabulación cruzada. Chi Cuadrado. Análisis Discriminante. Análisis Factorial. Análisis de Correspondencias. Análisis Clúster. Análisis Multidimensional no Métrico de Proximidades. Análisis Multidimensional no Métrico de Preferencias. AID. CHAID. Otros modelos.

3.3. Métodos docentes

La asignatura se desarrollará mediante la realización de diversas actividades: clases en el aula, tanto teóricas como prácticas, clases prácticas de laboratorio, trabajos individuales o de grupo, pruebas puntuables sobre ejercicios, tutorías individualizadas y examen final.

Clases en el aula: La teoría necesaria será expuesta por el profesor en clases en el aula. Se ilustrará su aplicación mediante ejemplos y ejercicios resueltos. Se realizarán ejercicios prácticos entregables con la supervisión del profesor.

Clases prácticas en el laboratorio: Los estudiantes realizarán prácticas de ordenador para familiarizarse con el manejo de los procedimientos estudiados en las clases de teoría. Se analizarán conjuntos de datos. Con el análisis y su interpretación se pretende que el estudiante sea capaz de segmentar de forma eficiente el mercado.

Se pondrán a disposición del alumno, a través de la plataforma educativa, documentos, ejercicios prácticos y enlaces con recursos. El detalle de los recursos a emplear, tanto en las clases como en el trabajo autónomo del alumno, se mostrará en los capítulos 4 y 5 de la presente memoria.

Tutorías. Concertadas en el despacho del profesor y tutorías vía correo electrónico.

3.4. Plan de trabajo

El 50% de las horas presenciales se desarrollarán en el aula, en las que se exponen todos los contenidos teóricos y prácticos del programa de la asignatura, se dan las indicaciones necesarias para facilitar la posterior labor de estudio de los alumnos y se proponen y resuelven ejercicios.

El 50% de las horas presenciales se desarrollarán en el laboratorio. En estas clases se implementa, sobre conjuntos de datos , todos y cada uno de los procedimientos desarrollados en las clases en el aula.

A lo largo del desarrollo de la asignatura se propondrá a los estudiantes la realización de trabajos evaluables y se realizará un examen final.

3.5. Evaluación

La evaluación continua del trabajo realizado por el estudiante en las clases se realizará mediante ejercicios entregables que representarán el 30% de la calificación final. Dentro de este porcentaje, cada trabajo evaluable pesará un 3%. Habrá un examen final de la asignatura que constituirá el 70% de la nota final.

La estructura del examen final consiste en resolver 3 o 4 ejercicios utilizando Jupyter Notebook. Se proporciona un notebook editable con los datos necesarios y los alumnos deben completar el resto del programa. Cada ejercicio se entrega en un notebook distinto, y una vez obtenidos los resultados, se deben registrar en un formulario que incluye preguntas rellenables (tanto de respuesta corta como larga) y preguntas tipo test con varias opciones. Al finalizar, el estudiante debe enviar tanto el formulario completo como los notebooks editados correspondientes a cada ejercicio.

Capítulo 4.
Recursos teóricos para la docencia

En esta sección se presentan materiales para apoyar la exposición de contenidos en el aula: presentaciones con los puntos más importantes de la teoría y soporte visual y una selección de fuentes documentales como bibliografía recomendada y artículos académicos. Estos recursos ofrecen un marco conceptual sólido y actualizado que facilita la planificación didáctica y promueve un aprendizaje significativo entre los estudiantes.

4.1. Presentación para el aula

Se muestran una serie de presentaciones en formato de diapositivas. El objetivo es proporcionar apoyo visual para la exposición de contenidos en el aula junto con la explicación que se aportaría como docente.

La presentación está dividida en dos grandes bloques:

- Las que se usarían en la explicación del concepto de segmentación en el contexto de marketing
- Las que se servirían de apoyo a la explicación de las técnicas estadísticas

Para cada uno de los bloques, se mostrarán las primeras diapositivas junto al guion correspondiente y, por motivos de limitación de extensión de la memoria, se continuará mostrando el resto de diapositivas sin guion.

Bloque 1: Concepto de segmentación de mercados

Tras el título de la asignatura, esta sería la primera diapositiva.

Comenzamos con el contexto histórico, nombrando el primer artículo en el que se habla propiamente de segmentar el mercado. En los 80's y 90's, el auge de la tecnología marcó un cambio en la forma de entender qué es la información y cómo se usa, por lo que se crearon bases de datos y software especializado. No fue sino hasta el siglo XXI que, debido a la

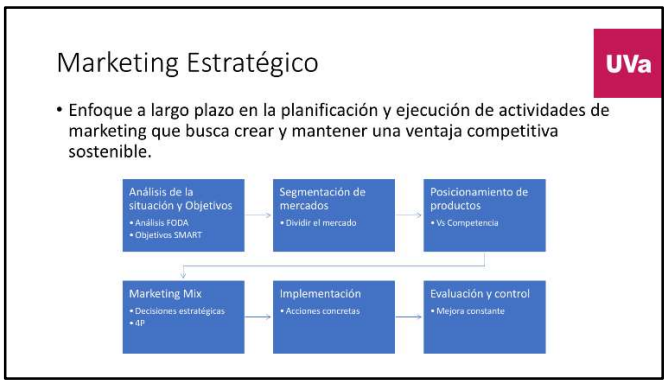
digitalización y a las redes sociales, la cantidad de información es tan masiva que es imprescindible el uso de técnicas estadísticas de segmentación de mercados.

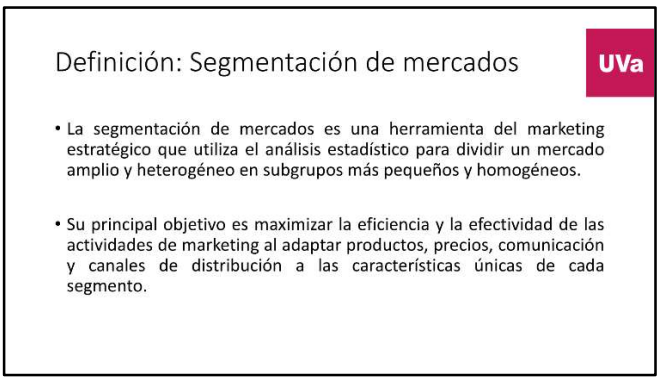
Como se puede leer, la investigación de mercado es la obtención sistemática de información para asistir a la dirección en la toma de decisiones comerciales de un problema concreto.

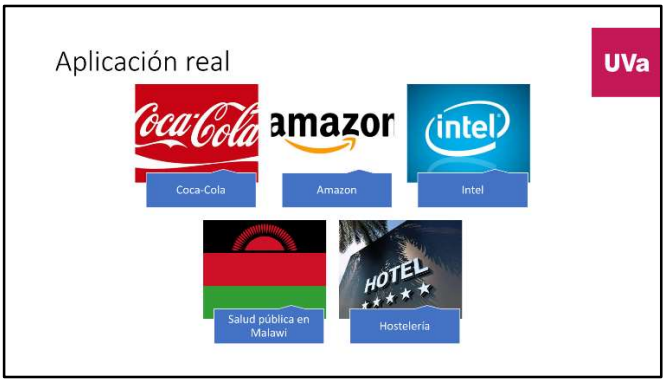
Es decir: la dirección de la empresa decide que quiere aumentar las ventas, por ejemplo, y se lo dice al departamento de marketing. Marketing aplica investigación de mercados y diseña un estudio: ¿qué necesitamos saber del mercado? ¿Qué de los clientes? ¿Qué de la competencia? ¿Qué del producto?... Acorde al problema de aumentar las ventas. A continuación, recogen toda la información necesaria y la analizan.

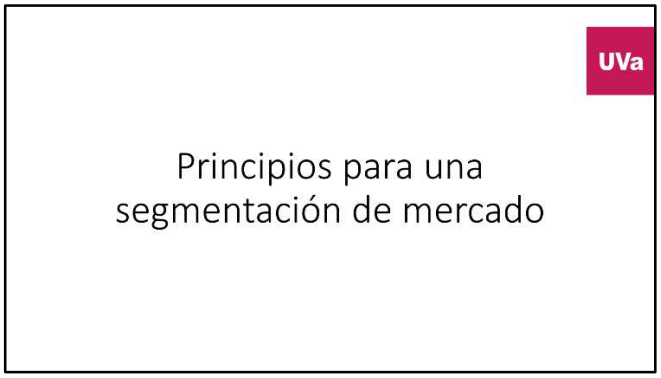
El resultado de dicho análisis se lo dicen a dirección, que son quienes tienen la decisión final.

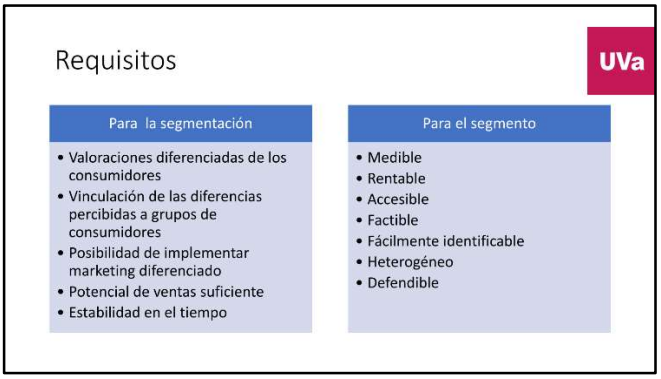
La presentación continuaría con las siguientes diapositivas:

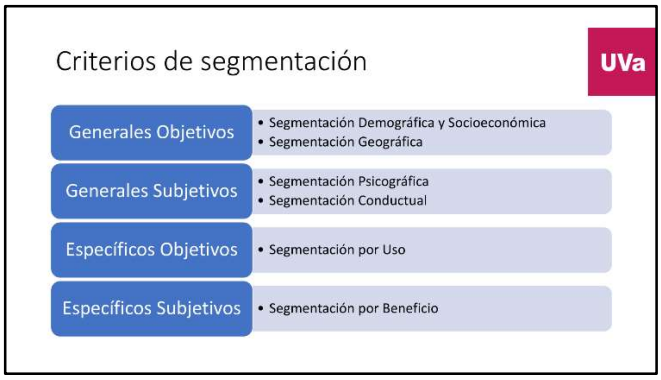


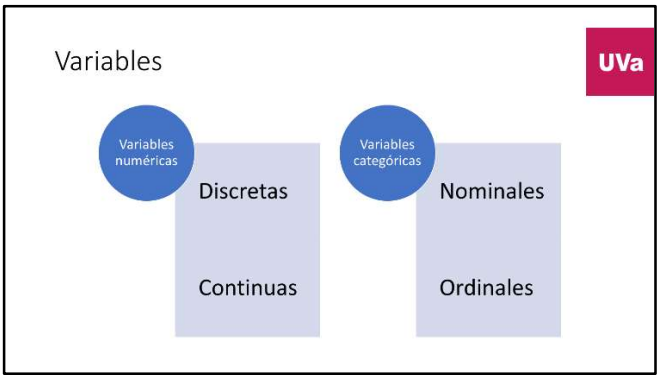


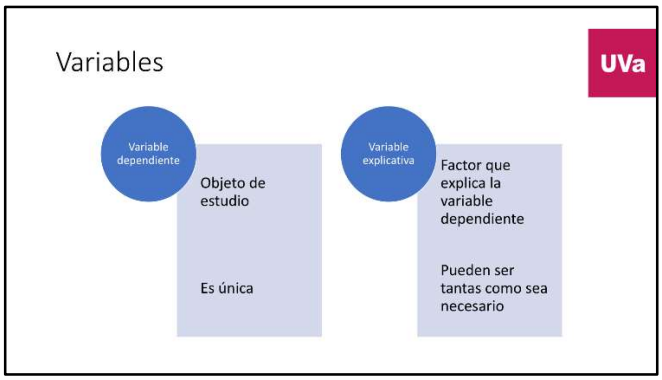


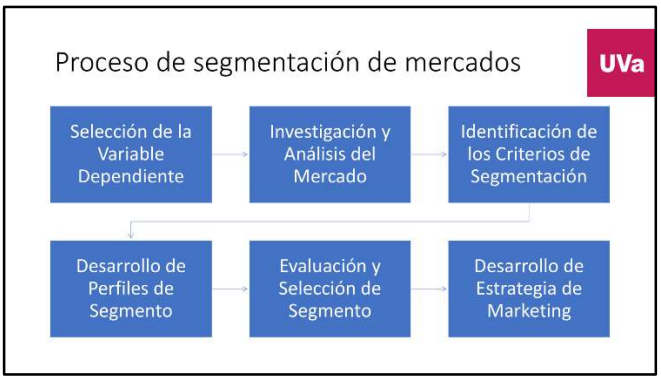






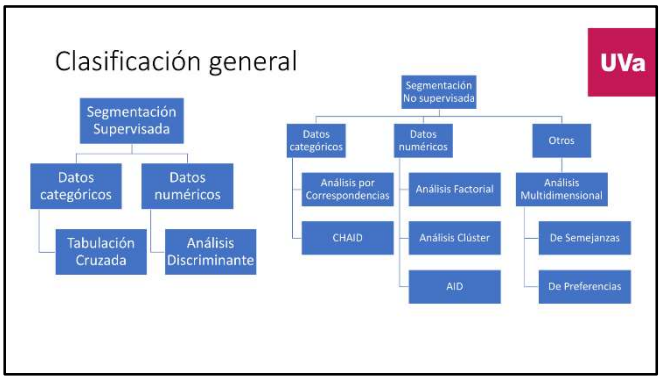






Bloque 2: Técnicas estadísticas de segmentación de mercado

Veremos ahora las técnicas que podemos aplicar para llevar a cabo el análisis del que hemos estado hablando. Para todas las técnicas, las dispositivas hacen referencia al ejercicio práctico que el alumno puede trabajar en el entorno Jupyter Notebook y que se describen en detalle en el capítulo de Recursos Prácticos de esta memoria.



Pero antes, vamos a situarnos. Hay un gran abanico de técnicas que nos pueden servir para segmentar el mercado y saber elegir la adecuada, según el tipo de estudio y el tipo de información que tenemos, es lo que va a lograr que obtengamos resultados más precisos, relevantes y accionables para tomar decisiones estratégicas acertadas.

Los dos grandes grupos de técnicas son la segmentación supervisada y la no supervisada.

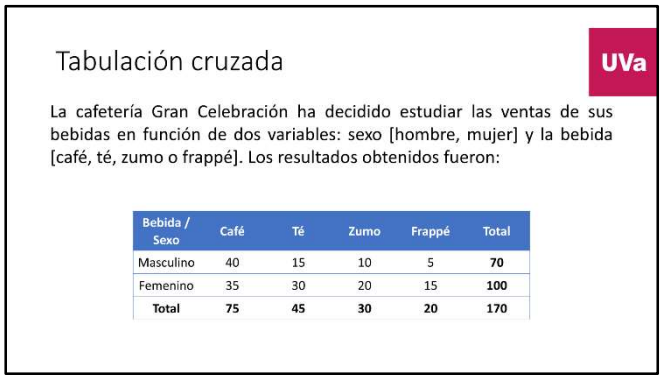
La segmentación supervisada es aquella que parte con los grupos ya creados. Es aquella en la que tenemos información previa sobre los propios segmentos y el objetivo es construir un modelo que nos ayude a clasificar correctamente nuevos individuos.

Cuando digo individuos pueden ser clientes, pero también pueden ser productos o empresas competencia.

En la segmentación no supervisada no tenemos ninguna información de los segmentos, por lo que vamos a ciegas. La explicaré con más profundidad más adelante.

Comenzaremos por la segmentación supervisada: aquella en la que ya conocemos los segmentos.

Y con datos categóricos, que, como ya explicamos, son datos que representan categorías o cualidades.



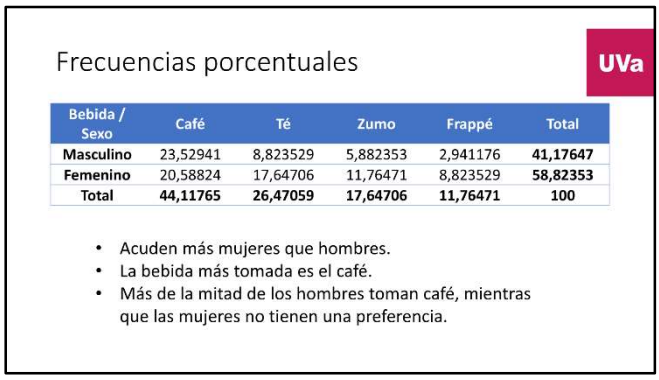
La técnica más sencilla para hacer segmentación de mercado es la tabulación cruzada. La podréis hacer a mano o con Excel y proporciona una clara relación entre las dos variables.

En esta técnica trabajaremos con una tabla de contingencia, o tabla cruzada, como la que veis en el ejemplo. Esta tabla es un resumen de las frecuencias de los pares de categorías. Es decir: un conteo. Y ahí también veis los totales.

Esta técnica es para dos variables categóricas, en el ejemplo: tipos de bebida y sexo; y en cada variable podéis tener tantas categorías como sea necesario, pero no infinito. Así que si una variable es numérica hay que discretizarla.

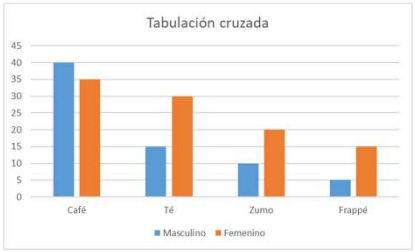
Veamos el ejemplo…

Se leería el enunciado y se explicaría cómo resolver la tabulación cruzada a mano. A continuación, muestro las diapositivas creadas para este propósito, simplificando la explicación que se daría en el aula:



Se explicaría el cálculo de las frecuencias porcentuales y qué análisis podemos extraer de ellas.

Gráficamente



Se explicaría el gráfico de barras del problema, confirmando el análisis anterior.

Test Chi-Cuadrado de independencia

- Tamaño de muestra suficientemente grande (frecuencias esperadas en cada celda de 5 o más)
- Observaciones independientes
- H0: Tipo de bebida y sexo del cliente no están relacionados.
- H1: Tipo de bebida y sexo del cliente sí están relacionados.

Bebida / Sexo	Café	Té	Zumo	Frappé
Masculino	40	15	10	5
Femenino	35	30	20	15

Debido a que la tabulación cruzada no incorpora ningún tipo de contraste para estudiar si las diferencias entre las variables con significativas o no, se explica el test Chi-cuadrado de independencia y se continua con el ejemplo.

Frecuencias esperadas

$$f_i(i,j) = \frac{R_i \times C_j}{N}$$

Bebida / Sexo	Café	Té	Zumo	Frappé	Total
Masculino	30,88235	18,52941	12,35294	8,235294	**70**
Femenino	44,11765	26,47059	17,64706	11,76471	**100**
Total	**75**	**45**	**30**	**20**	**170**

Para el cálculo del test de Chi-cuadrado de independencia, se explica el cálculo de las frecuencias esperadas.

Estadístico Chi-cuadrado

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{Eij}$$

$\chi^2 = 8{,}641666667$

Valor Crítico

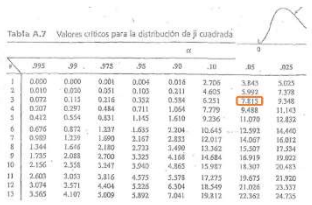


Tabla A.7 Valores críticos para la distribución de χ^2 cuadrada

Grados de libertad

$$gl = (i-1) \times (j-1) = 3$$

Y se explica la obtención del estadístico Chi-cuadrado y la forma de comprobar, con él, si la diferencia entre las variables estudiadas es significativa usando la tabla de los valores críticos de la distribución Chi-cuadrado.

P-valor

Es un valor entre 0 y 1 que indica la probabilidad de que la hipótesis nula es verdadera.

- Si el p-valor es menor que 0.05, rechazamos la hipótesis nula.
- Si el p-valor es mayor que 0.05, decimos que no se puede rechazar la hipótesis nula, pues no hay diferencias significativas.

Ejercicio: TCruzada1.ipynb

Finalmente, se explica el p-valor y su importancia. Para su cálculo se acudirá a la plataforma educativa que los alumnos usarán de aquí en adelante para entregar los ejercicios y comunicarse con el profesor y se volverá a ver el ejercicio de ejemplo, ahora en Python, en un Jupyter notebook.

De esta forma obtendríamos el p-valor y lo interpretaríamos y debe ser acorde a lo ya analizado previamente de forma manual.

Para el resto de técnicas, se muestran exclusivamente las dispositivas sin guion:

Segmentación Supervisada

Datos numéricos

LDA

Análisis Discriminante Lineal

- Función discriminante para una clase k :

$$\delta k(x)=x^T*\Sigma-1\mu k-21\mu k^T\neg*\Sigma-1\mu k+\log(\pi k)$$

- Suposiciones:
  - Normalidad Multivariante
  - Igualdad de Covarianzas
  - Linealidad

Ejercicio: LDA1.ipynb

QDA

Análisis Discriminante Cuadrático

- Función discriminante para una clase k :

$$\delta k(x)=-21\log|\Sigma k|-21(x-\mu k)T\Sigma k-1(x-\mu k)+\log(\pi k)$$

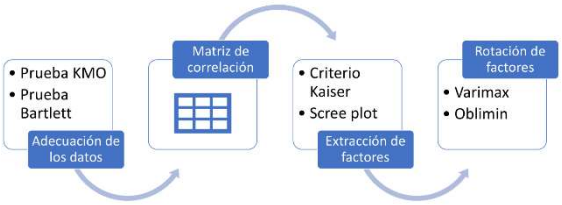
- Suposiciones:
  - Normalidad Multivariante
- Riesgo de sobreajuste:
  - Nº datos = 10 x p_variables x k_clases

Ejercicio: QDA1.ipynb

Segmentación No Supervisada

Análisis Factorial

Objetivo: Reducir la dimensionalidad.



Análisis Factorial

Una empresa de telecomunicaciones busca segmentar a sus usuarios en base a comportamientos observados y no sus preferencias personales.

Usuario	Minutos al mes	Número de llamadas	Datos móviles (GB)
U1	800	200	3.5
U2	750	180	4.0
U3	300	50	12.0
U4	280	45	11.0
U5	700	190	2.0

Ejercicio: Factorial1.ipynb

CA

Análisis por Correspondencias

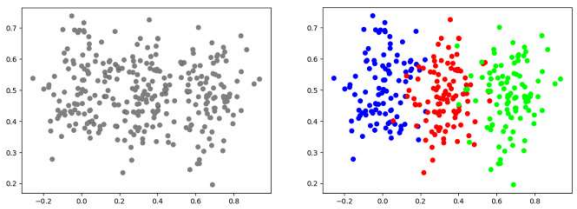
Es una ampliación de la tabulación cruzada con técnicas del análisis factorial.

Objetivos:

- Reducir la dimensionalidad
- Representar gráficamente las relaciones entre categorías de las variables
- Identificar asociaciones y patrones

Ejercicio: Corr1.ipynb

Análisis Clúster



Tipos de clustering

Clustering Particional	Clustering Jerárquico	Basados en Densidad	Basados en Distribución
• K-Means	• Linkage Method: - Ward	• DBSCAN	• GMM

K-Means

- Divide los datos en k clústeres.
- Minimiza la suma de las distancias cuadradas entre los puntos y el centroide del clúster.

- Es rápido y fácil de implementar.
- Requiere especificar el número de clústeres k.
- Puede converger a un óptimo local.

Ejercicio: Kmeans1.ipynb

Ward

- Es un método de enlace.
- Produce un dendrograma que muestra la fusión o división de clústeres.
- No requiere el número de clústeres de antemano.

Ejercicio: Ward1.ipynb

MDS

Análisis Multidimensional

- Datos subjetivos
- Buscamos representar visualmente las percepciones de los encuestados

MDS métrico	MDS no métrico
• Disimilitudes numéricas • Conservar distancias reales	• Orden o ranking de disimilitudes • No se conservan proporciones exactas

Semejanza

- Los datos son pares de productos, marcas o atributos.

Hicimos una encuesta a consumidores, y obtuvimos la siguiente matriz de disimilitudes:

Escala 1–9, donde
1 = muy parecidos
9 = muy distintos

Ejercicio: Semejanza1.ipynb

	A	B	C	D
A	0	3	7	6
B	3	0	5	4
C	7	5	0	2
D	6	4	2	0

Preferencia

- Los datos son preferencias individuales respecto a un conjunto de productos, marcas o atributos.

3 consumidores califican 4 productos (A, B, C, D) :

Escala de 1 a 10
1 = menor preferencia
10 = mayor preferencia

Ejercicio: Preferencia1.ipynb

Producto	Persona 1	Persona 2	Persona 3
A	8	7	6
B	6	5	4
C	4	8	7
D	2	3	5

AID

Automatic Interaction Detection

- Crea un árbol de decisión
- Parte de una variable dependiente numérica
- Utiliza ANOVA para hacer las divisiones
  - Divisiones binarias
  - Maximiza la diferencia en promedios:
- Previsión de gasto de un cliente.

CHAID

Chi-Squared Automatic Interaction Detection

- Crea un árbol de decisión
- Parte de una variable dependiente categórica
- Utiliza test Chi-cuadrado para hacer las divisiones
  - Permite divisiones múltiples
  - Segmentos significativamente distintos:
- Segmentación de mercado.
- Análisis de comportamiento.

Ejercicio: CHAID.ipynb

Otros modelos

Por último, tras todas las técnicas explicadas y todos los ejercicios, se nombrarían las diferentes ramas de la segmentación de mercados con técnicas más avanzadas que las vistas, técnicas con enfoques diferentes o técnicas mixtas

De esta forma los alumnos tendrían una visión más amplia y completa de la segmentación de mercado.

Todas las diapositivas se encuentran en el Anexo I.

4.2. Referencias propuestas para el alumno

A continuación, se ofrece una selección de fuentes teóricas y materiales de consulta. El objetivo de todas ellas es enriquecer el abordaje de los contenidos. Incluyen bibliografía recomendada y artículos académicos que proporcionan marcos conceptuales actualizados diversas perspectivas sobre los temas tratados.

Bibliografía recomendada

Esta es una selección de lecturas recomendadas diferenciadas entre el conocimiento de segmentación de mercados y el de las técnicas estadísticas.

- **Sobre segmentación de mercados:**

  - Cruz Roche, I. (s.f.). *Fundamentos de marketing*. Editorial Ariel. (Roche, 1991)
  - Grande Esteban, I. & Abascal Fernández, E. (s.f.). *Fundamentos y técnicas de investigación comercial*. Editorial ESIC. (Grande Esteban, 2017)
  - Ortega Martínez, E. (s.f.). *Manual de Investigación Comercial*. Editorial Pirámide. (Ortega Martínez, 1992)
  - Santesmases Mestre, M. (s.f.). *Marketing: conceptos y estrategias*. Editorial Pirámide. (Santesmases Mestre, 1996)

- **Sobre técnicas estadísticas:**

  - Flury, B. (1997). *A first Course in Multivariate Statistics*. Editorial Springer. (Flury, 1997)
  - Hastie, T. & Tibishirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. Editorial Springer. (Hastie, Tibshirani, & Friedman, 2001)
  - Ortega Martínez, E. (s.f.). *Manual de Investigación Comercial*. Editorial Pirámide. (Ortega Martínez, 1992)
  - Peña, D. (2002). *Análisis de Datos Multivariantes*. Editorial Mc Graw Hill. (Peña, 2002)
  - Seber, G.A.F. (1984). *Multivariate Observations*. Editorial Wiley. (Seber, 1984)
  - Srivastava, M.S. (2002). *Methods of Multivariate Statistics*. Editorial Wiley. (Srivastava, 2002)

Artículos

Además de la bibliografía principal, se han seleccionado algunos trabajos académicos que abordan temáticas relacionadas con los contenidos de la asignatura. Estas investigaciones ofrecen enfoques actuales, aplicados y contextualizados:

- Escobar, M. (1998). *Las aplicaciones del análisis de segmentación: El procedimiento CHAID*. Universidad de Salamanca. (Escobar, 1998)
- Fuente Fernández, S. (2011). *Análisis correspondencias simples y múltiples*. Universidad Autónoma de Madrid. (De la Fuente Fernández, 2011)
- Lozano Ruiz, I. *Optimización de campañas de marketing en un centro de comercio mediante segmentación de clientes.* (Lozano Ruiz, 2025)
- Palacios, J. E. R., & Rivera, J. R. L. (2025). *Preferencias y hábitos de consumo del mercado de café tostado en el Distrito Central, Francisco Morazán.* (Rodríguez Palacios & López Rivera, 2025)

Capítulo 5.
Recursos prácticos para la docencia

Con el objetivo de complementar el marco teórico y favorecer un aprendizaje activo, esta sección reúne diversos recursos prácticos diseñados para fortalecer las habilidades y competencias de los estudiantes. Se incluyen problemas resueltos que ejemplifican la aplicación de conceptos, ejercicios propuestos para estimular el análisis y la resolución autónoma, así como problemas de autoevaluación que permiten a los alumnos medir su progreso. Además, se presentan materiales de evaluación que facilitan la valoración objetiva del aprendizaje.

5.1. Problemas resueltos

Esta sección presenta una colección de problemas desarrollados paso a paso, con el objetivo de ilustrar la aplicación práctica de los contenidos teóricos. Debido a la extensión de esta memoria se verá un ejercicio resuelto completo, estando en el Anexo II el conjunto de ejercicios que abarcan todas las técnicas estadísticas vistas.

El ejercicio seleccionado es aquel en el que se ilustra el Análisis Clúster usando la técnica K-means.

Se ha elegido este problema por su complejidad y representatividad, ya que permite demostrar los pasos del método K-means y diferenciarse de las demás técnicas estadísticas explicadas. Además, en este ejemplo se pone de manifiesto que datos aparentemente no diferenciados por grupos, en cuyo gráfico de dispersión se observa una gran nube de datos; con el método adecuado, pueden ser segmentados.

Análisis Clúster - K-means

Ejercicio Resuelto

Un centro comercial quiere segmentar a sus clientes para ofrecer promociones más personalizadas. Para ello, ha recolectado datos de 200 clientes. Tu tarea es aplicar K-Means para identificar patrones de comportamiento y dividir a los clientes en grupos significativos.

1.0 Datos

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

np.random.seed(321)

# Datos
n = 200
ingreso = np.random.normal(3000, 1000, n)       # en euros
puntaje_gasto = np.random.randint(1, 101, n)    # entre 1 y 100
edad = np.random.randint(18, 65, n)

df = pd.DataFrame({
    'Ingreso': ingreso,
    'PuntajeGasto': puntaje_gasto,
    'Edad': edad
})

df.head()
```

	Ingreso	PuntajeGasto	Edad
0	3172.519469	2	41
1	4635.482532	58	59
2	3037.336403	55	36
3	2115.850315	81	31
4	1856.807744	48	61

1.1 Estandarización

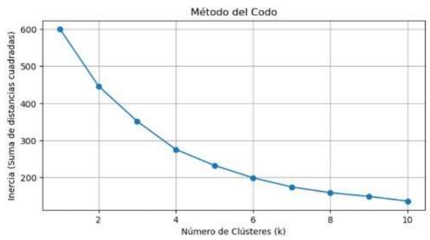
```python
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df)
```

1.2 Gráfico de codo

```python
inertias = []
K_range = range(1, 11)

for k in K_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(X_scaled)
    inertias.append(kmeans.inertia_)

plt.figure(figsize=(8, 4))
plt.plot(K_range, inertias, marker='o')
plt.title("Método del Codo")
plt.xlabel("Número de Clústeres (k)")
plt.ylabel("Inercia (Suma de distancias cuadradas)")
plt.grid(True)
plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
```

```
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```

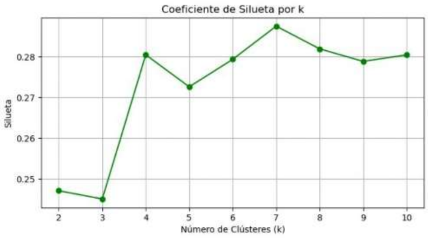


1.3 Coeficiente silueta

```python
# □ Coeficiente de silueta para distintos k
silhouette_scores = []
K_range_eval = range(2, 11)

for k in K_range_eval:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    labels = kmeans.fit_predict(X_scaled)
    score = silhouette_score(X_scaled, labels)
    silhouette_scores.append(score)

plt.figure(figsize=(8, 4))
plt.plot(K_range_eval, silhouette_scores, marker='o', color='green')
plt.title("Coeficiente de Silueta por k")
plt.xlabel("Número de Clústeres (k)")
plt.ylabel("Silueta")
plt.grid(True)
plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```



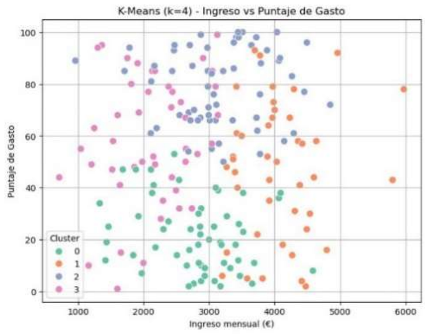
1.4 K-means final

```
# Elegimos k=4
k_final = 4
kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=10)
df['Cluster'] = kmeans_final.fit_predict(X_scaled)
```

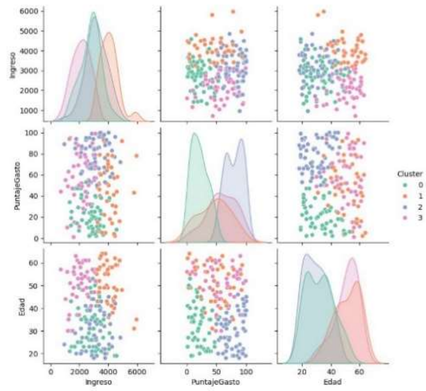
```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```

1.5 Gráfico de dispersión

```
# Ingreso vs. Puntaje de Gasto
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='Ingreso', y='PuntajeGasto', hue='Cluster',
palette='Set2', s=70)
plt.title(f"K-Means (k={k_final}) - Ingreso vs Puntaje de Gasto")
plt.xlabel("Ingreso mensual (€)")
plt.ylabel("Puntaje de Gasto")
plt.legend(title="Cluster")
plt.grid(True)
plt.show()
```

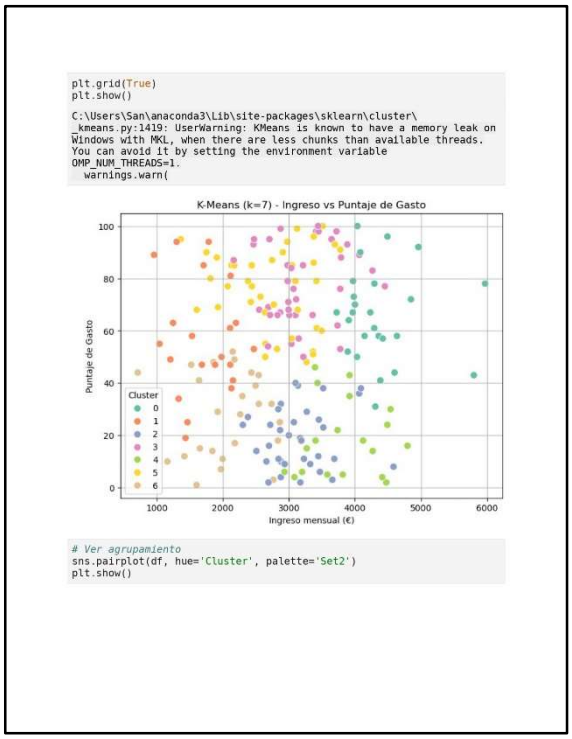



```
# Ver agrupamiento
sns.pairplot(df, hue='Cluster', palette='Set2')
plt.show()
```



1.6 Repetimos con k = 7

```
# Elegimos k=7
k_final = 7
kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=10)
df['Cluster'] = kmeans_final.fit_predict(X_scaled)

# Ingreso vs. Puntaje de Gasto
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='Ingreso', y='PuntajeGasto', hue='Cluster',
palette='Set2', s=70)
plt.title(f"K-Means (k={k_final}) - Ingreso vs Puntaje de Gasto")
plt.xlabel("Ingreso mensual (€)")
plt.ylabel("Puntaje de Gasto")
plt.legend(title="Cluster")
```

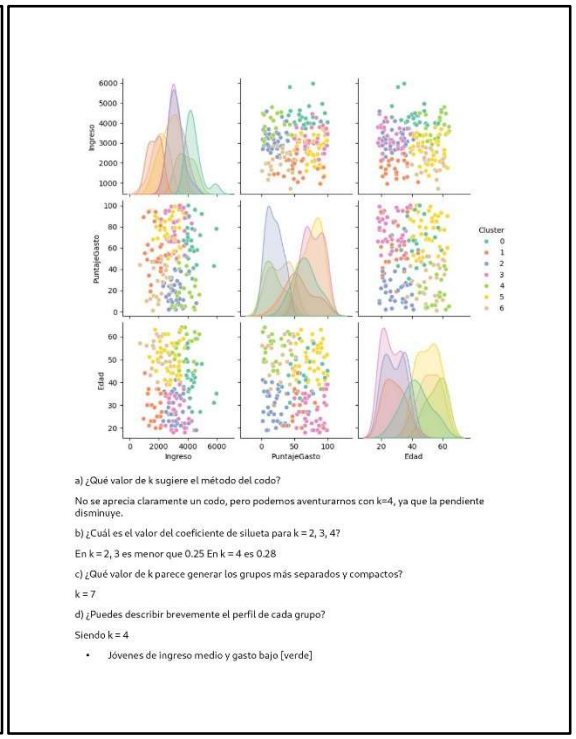


```
plt.grid(True)
plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```



```
# Ver agrupamiento
sns.pairplot(df, hue='Cluster', palette='Set2')
plt.show()
```



a) ¿Qué valor de k sugiere el método del codo?

No se aprecia claramente un codo, pero podemos aventurarnos con k=4, ya que la pendiente disminuye.

b) ¿Cuál es el valor del coeficiente de silueta para k = 2, 3, 4?

En k = 2, 3 es menor que 0.25 En k = 4 es 0.28

c) ¿Qué valor de k parece generar los grupos más separados y compactos?

k = 7

d) ¿Puedes describir brevemente el perfil de cada grupo?

Siendo k = 4

- Jóvenes de ingreso medio y gasto bajo [verde]
- Jóvenes de ingreso medio y gasto alto [azul]
- Mayores de ingreso alto y gasto medio [rojo]
- Mayores de ingreso bajo y gasto medio [rosa]

e) ¿Qué tipo de estrategia de marketing aplicarías para cada segmento?

Siguiendo el orden de la pregunta anterior:

- Ofertas de entrada, campañas de fidelización y muostrar el valor y beneficios concretos del producto.
- Campañas experienciales, exclusividad asequible, personalización del producto o servicio, up-selling y cross-selling.
- Servicio premium, email marketing segmentado y contenido informativo, productos complementarios de valor añadido y destacar beneficios a largo plazo, garantías y prestigio de marca.
- Promociones por ahorro o conveniencia, facilidades de pago y programas de fidelidad.

f) ¿Qué clúster representa a los clientes más valiosos para la empresa?

Los jóvenes de ingreso medio y gasto alto, pues tienen una mayor proyección y se podría esperar que con el paso del tiempo sus ingresos aumentaran como lo haría también su gasto.

g) ¿Cómo podrías usar esta segmentación para personalizar promociones?

Se podrían usar las estrategias de marketing que ya se explicaron anteriormente y llevarlas a cabo coordinadamente con los eventos anuales como pueden ser: el inicio de las vacaciones, la navidad y la temporada de rebajas.

h) ¿Los clústeres se visualizan claramente en un gráfico de dispersión?

No

i) ¿Se superponen los clústeres en algunas combinaciones de variables?

Sí, en el punto medio de las tres variables hay clientes que podrían pertenecer a cualquiera de las clases.

Cada problema en el anexo incluye el enunciado, los datos iniciales, la solución paso a paso y unas preguntas finales, junto con análisis de resultados. Todos los ejercicios, uno por técnica explicada en esta memoria, están en el Anexo II.

5.2. Problemas propuestos evaluables

A continuación, se ofrecen ejercicios adicionales para que los estudiantes practiquen de forma autónoma. Estos problemas están diseñados para reforzar los conocimientos adquiridos durante las horas de clase. Los

ejercicios varían en dificultad: algunos son similares a los resueltos previamente y otros presentan un mayor nivel de complejidad. Esto permite a los estudiantes identificar errores, medir su progreso y consolidar los aprendizajes. Todo ello alojado en la plataforma didáctica que se explica en el siguiente capítulo de esta memoria.

Elementos para la evaluación de los problemas

Para facilitar el trabajo individual de los estudiantes y asegurar que se practiquen los contenidos de forma activa, se han definido dos métodos complementarios relacionados con los problemas propuestos. Ambos métodos están pensados para que el docente pueda obtener una nota de evaluación continua objetiva y que, al mismo tiempo, pueda estar informado de qué parte de la materia impartida necesita mayor tiempo de trabajo en el aula.

Los dos métodos son:

1. Cuadernos Jupyter entregables

A medida que se avance en la teoría, el docente podrá ir compartiendo cuadernos Jupyter (notebook) de las diferentes técnicas estudiadas con los alumnos. Con cada una de las técnicas habrá, además de ejercicios resueltos, un ejercicio en un cuaderno Jupyter editable que deberá ser entregado para su evaluación. Dicha entrega tendrá como fecha de finalización dos o tres días de haber visto la técnica de segmentación en el aula.

Este método permite:

- Comprobar la correcta aplicación práctica de los contenidos.
- Valorar tanto los resultados como la claridad del desarrollo del código y razonamiento.
- Evaluar la capacidad de documentar y justificar adecuadamente los procedimientos utilizados.

El cuaderno Jupyter puede incluir:

- Celdas de código con la solución de cada ejercicio.
- Comentarios o explicaciones escritas.
- Visualizaciones (si aplica).

Ejemplo de un cuaderno Jupyter entregable desde el punto de vista del alumno:

Guía para el alumno

Este cuaderno contiene las librerías y datos iniciales. Resuelve cada paso siguiendo las indicaciones. Añade tus propias celdas de código debajo de cada sección.

2.0 Datos

In [21]:
```
import pandas as pd
import numpy as np
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from matplotlib.colors import ListedColormap

np.random.seed(101)

# Segmento: Alta fidelidad
alta = pd.DataFrame({
    'FrecuenciaCompra': np.random.normal(12, 2, 30),
    'DescuentoMedio': np.random.normal(5, 2, 30),
    'DiasDesdeUltima': np.random.normal(5, 2, 30),
    'ValorTicketMedio': np.random.normal(60, 5, 30),
    'Segmento': 'Alta fidelidad'
})

# Segmento: Cazador de ofertas
ofertas = pd.DataFrame({
    'FrecuenciaCompra': np.random.normal(6, 1.5, 30),
    'DescuentoMedio': np.random.normal(20, 3, 30),
    'DiasDesdeUltima': np.random.normal(15, 3, 30),
    'ValorTicketMedio': np.random.normal(45, 5, 30),
    'Segmento': 'Cazador de ofertas'
})

# Segmento: Ocasional
ocasional = pd.DataFrame({
    'FrecuenciaCompra': np.random.normal(3, 1, 30),
    'DescuentoMedio': np.random.normal(10, 4, 30),
    'DiasDesdeUltima': np.random.normal(30, 5, 30),
    'ValorTicketMedio': np.random.normal(40, 7, 30),
    'Segmento': 'Ocasional'
})

# Unir los datos
df = pd.concat([alta, ofertas, ocasional], ignore_index=True)
df = df.sample(frac=1, random_state=42).reset_index(drop=True).round(2)

df.head()
```

Out[21]:

	FrecuenciaCompra	DescuentoMedio	DiasDesdeUltima	ValorTicketMedio	Segmento
0	7.09	20.28	11.29	43.51	Cazador de ofertas
1	8.59	5.29	4.67	54.60	Alta fidelidad
2	5.39	16.81	15.50	40.03	Cazador de ofertas
3	2.20	5.65	29.74	44.52	Ocasional
4	17.41	6.28	5.77	57.16	Alta fidelidad

Paso 1: Análisis exploratorio

Describe la estructura del dataset, tipos de variables, valores nulos, etc.

In [ ]:
```
# Tu código aquí
```

Paso 2: Visualización

Representa gráficamente las variables para entender sus relaciones.

In [ ]:
```
# Tu código aquí
```

Paso 3: Preparación de datos

Codifica, escala o transforma las variables si es necesario.

In [ ]:
```
# Tu código aquí
```

Paso 4: División del dataset

Separa los datos en conjunto de entrenamiento y test.

In [ ]:
```
# Tu código aquí
```

Paso 5: Entrenamiento del modelo LDA

Aplica LDA con los datos preparados.

In [ ]:
```
# Tu código aquí
```

Paso 6: Evaluación del modelo

Calcula métricas e interpreta los resultados.

In [ ]:
```
# Tu código aquí
```

Paso 7: Conclusión

Resume tus hallazgos e interpreta el valor del modelo.

Ilustración 3. Ejemplo de cuaderno Jupyter entregable

2. Formulario con preguntas sobre los ejercicios

Junto al cuaderno se proporciona un formulario con preguntas sobre los problemas propuestos que se ha de rellenar y enviar a la vez que el cuaderno Jupyter entregable al que está asociado. En dicho formulario consta de preguntas acerca del ejercicio y de la comprensión del mismo.

Este formulario permite:

- Valorar la comprensión conceptual de los ejercicios realizados.
- Detectar si el estudiante ha trabajado personalmente en la resolución.
- Facilitar una corrección automatizada.
- Obtención de estadísticas útiles para el docente.

Las preguntas pueden ser de:

- Opción múltiple
- Respuesta corta
- Preguntas de reflexión o explicación breve

Ejemplo de un formulario desde el punto de vista de un alumno:

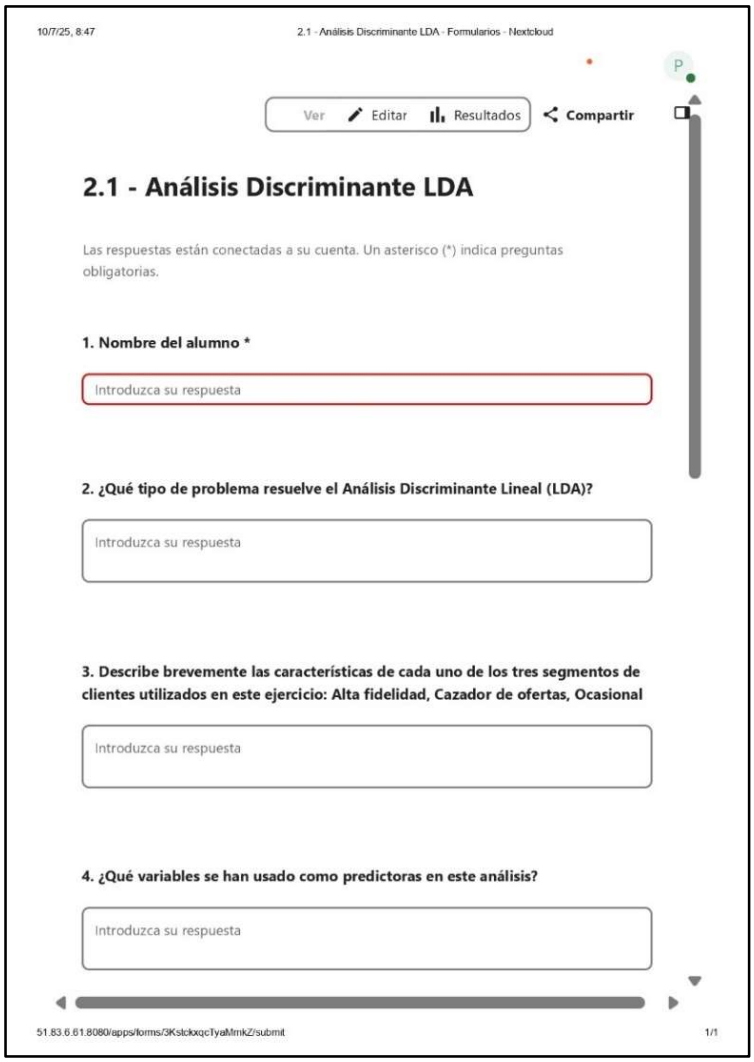


Ilustración 4. Ejemplo de formulario

Objetivo de usar ambos métodos combinados

Se busca asegurar un aprendizaje activo y profundo de la materia, practicando con diferentes variables y problemas las técnicas explicadas. Evaluar tanto la parte práctica (resolución de ejercicios con el cuaderno de Jupyter) como la teórica y reflexiva (mediante el formulario) hace que la evaluación continua por parte del docente sea más objetiva y sencilla, ya que le permite tener información detallada de cómo van aprendiendo los alumnos.

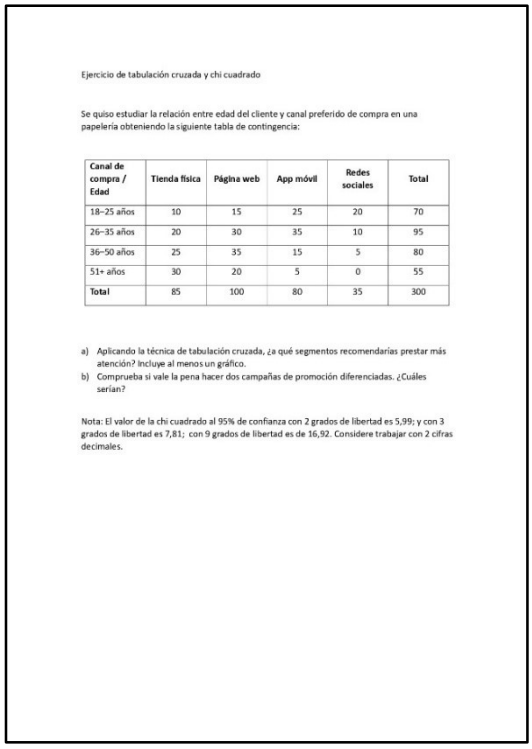
Además, el hecho de que estos dos métodos sean ejercicios para hacer fuera del aula fomenta la autonomía del estudiante.

5.3. Examen final

Finalmente, se propone un instrumento de evaluación en formato de examen, que permite valorar de manera integral el nivel de comprensión y dominio de los contenidos abordados. Dicho examen estará compuesto por varios ejercicios similares a los vistos anteriormente. Cada ejercicio deberá ser de una técnica distinta para lograr abarcar el mayor número de competencias y contenidos claves, asegurando así una evaluación equilibrada y representativa del aprendizaje esperado.

A continuación, se detallará un ejemplo de examen con: selección de ejercicios, solución de los mismos y una guía de corrección en la que se detallará la ponderación de cada apartado.

El examen propuesto consta de cuatro ejercicios. Detallaré a continuación la selección de los mismos:



Ejercicio de tabulación cruzada y chi cuadrado

Se quiso estudiar la relación entre edad del cliente y canal preferido de compra en una papelería obteniendo la siguiente tabla de contingencia:

Canal de compra / Edad	Tienda física	Página web	App móvil	Redes sociales	Total
18–25 años	10	15	25	20	70
26–35 años	20	30	35	10	95
36–50 años	25	35	15	5	80
51+ años	30	20	5	0	55
Total	85	100	80	35	300

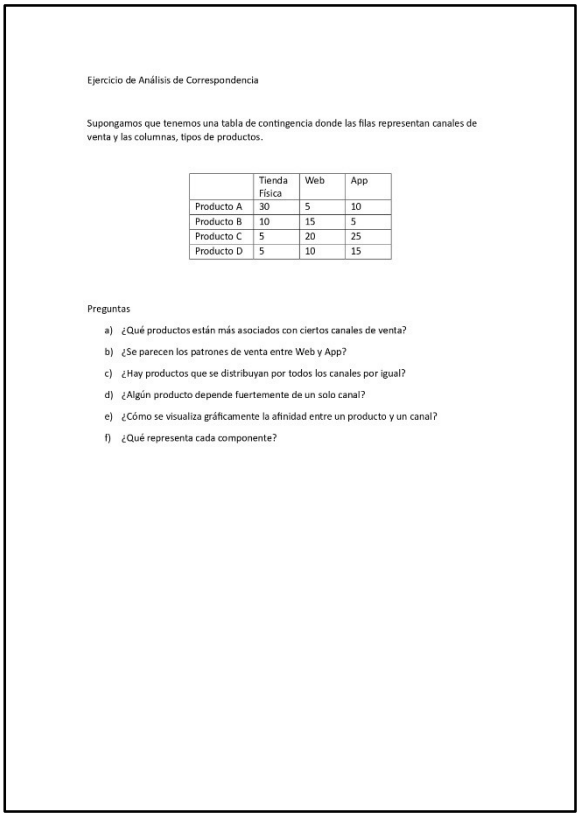
a) Aplicando la técnica de tabulación cruzada, ¿a qué segmentos recomendarías prestar más atención? Incluye al menos un gráfico.
b) Comprueba si vale la pena hacer dos campañas de promoción diferenciadas. ¿Cuáles serían?

Nota: El valor de la chi cuadrado al 95% de confianza con 2 grados de libertad es 5,99; y con 3 grados de libertad es 7,81; con 9 grados de libertad es de 16,92. Considere trabajar con 2 cifras decimales.

Tabulación Cruzada y chi cuadrado.

Es la técnica básica y esencial en estudios de mercado, ya que permite identificar rápidamente asociaciones y diferencias entre grupos y es base para una segmentación descriptiva inicial.

Permite que los alumnos trabajen con ordenador o a mano, elección que queda a cargo del docente. En ambos casos, se evalúa la capacidad de analizar relaciones entre variables categóricas y la comprensión de patrones de comportamiento en diferentes segmentos de consumidores.



Ejercicio de Análisis de Correspondencia

Supongamos que tenemos una tabla de contingencia donde las filas representan canales de venta y las columnas, tipos de productos.

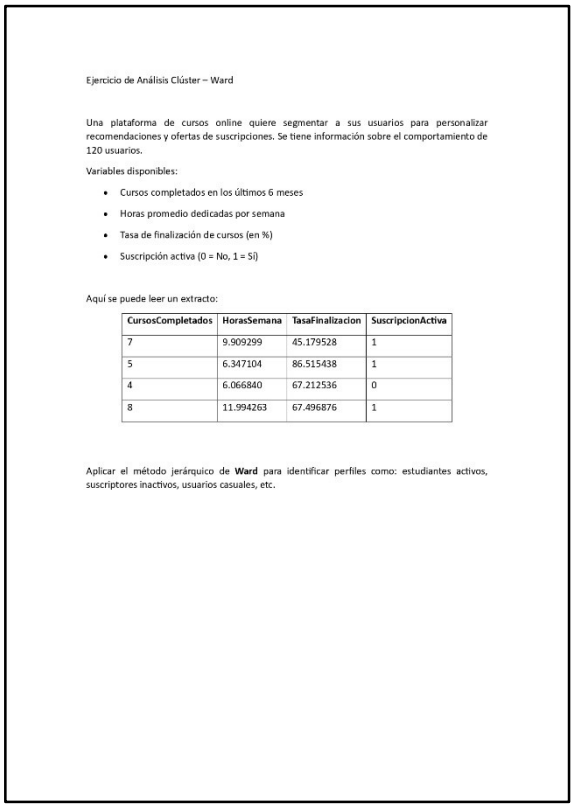
	Tienda Física	Web	App
Producto A	30	5	10
Producto B	10	15	5
Producto C	5	20	25
Producto D	5	10	15

Preguntas

a) ¿Qué productos están más asociados con ciertos canales de venta?
b) ¿Se parecen los patrones de venta entre Web y App?
c) ¿Hay productos que se distribuyan por todos los canales por igual?
d) ¿Algún producto depende fuertemente de un solo canal?
e) ¿Cómo se visualiza gráficamente la afinidad entre un producto y un canal?
f) ¿Qué representa cada componente?

Análisis de Correspondencias

El análisis de correspondencias permite interpretar resultados complejos de manera intuitiva y visual. Está pensado para evaluar la interpretación gráfica de asociaciones entre variables (marcas, atributos, preferencias) y para analizar la carga de las componentes.



Ejercicio de Análisis Clúster – Ward

Una plataforma de cursos online quiere segmentar a sus usuarios para personalizar recomendaciones y ofertas de suscripciones. Se tiene información sobre el comportamiento de 120 usuarios.

Variables disponibles:

- Cursos completados en los últimos 6 meses
- Horas promedio dedicadas por semana
- Tasa de finalización de cursos (en %)
- Suscripción activa (0 = No, 1 = Si)

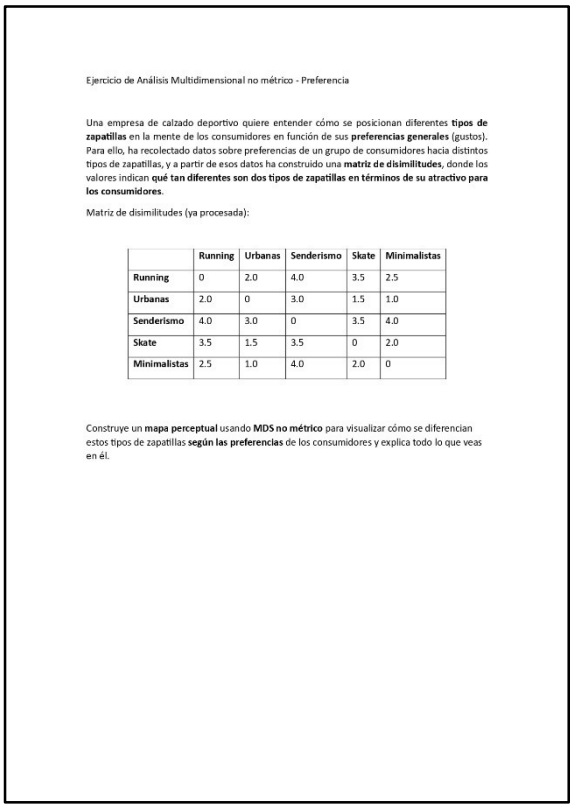
Aquí se puede leer un extracto:

CursosCompletados	HorasSemana	TasaFinalizacion	SuscripcionActiva
7	9.909299	45.179528	1
5	6.347104	86.515438	1
4	6.066840	67.212536	0
8	11.994263	67.496876	1

Aplicar el método jerárquico de **Ward** para identificar perfiles como: estudiantes activos, suscriptores inactivos, usuarios casuales, etc.

Modelo Ward

El análisis de clúster y específicamente el método de Ward evalúa si el estudiante puede identificar segmentos reales basados en múltiples variables cuantitativas o cualitativas. Además, se comprueba la habilidad para interpretar dendrogramas, podarlo y definir grupos homogéneos.



Ejercicio de Análisis Multidimensional no métrico - Preferencia

Una empresa de calzado deportivo quiere entender cómo se posicionan diferentes **tipos de zapatillas** en la mente de los consumidores en función de sus **preferencias generales** (gustos). Para ello, ha recolectado datos sobre preferencias de un grupo de consumidores hacia distintos tipos de zapatillas, y a partir de esos datos ha construido una **matriz de disimilitudes**, donde los valores indican **qué tan diferentes son dos tipos de zapatillas en términos de su atractivo para los consumidores**.

Matriz de disimilitudes (ya procesada):

	Running	Urbanas	Senderismo	Skate	Minimalistas
Running	0	2.0	4.0	3.5	2.5
Urbanas	2.0	0	3.0	1.5	1.0
Senderismo	4.0	3.0	0	3.5	4.0
Skate	3.5	1.5	3.5	0	2.0
Minimalistas	2.5	1.0	4.0	2.0	0

Construye un **mapa perceptual** usando **MDS no métrico** para visualizar cómo se diferencian estos tipos de zapatillas **según las preferencias** de los consumidores y explica todo lo que veas en él.

Análisis Multidimensional no Métrico

Por último, un análisis multidimensional no métrico. En este caso, se ha elegido las preferencias. Incluir este ejercicio permite evaluar la orientación práctica hacia el diseño de productos o estrategias de posicionamiento, ya que gran parte de las decisiones de compra son subjetivas. También pone de manifiesto la capacidad para entender y representar el mapa perceptual.

Solución

La solución de los cuatro ejercicios se incluye en el Anexo III, junto con sus respectivos enunciados, a modo de ejercicios resueltos. Dado que ya se ha presentado un ejemplo en el cuerpo de esta memoria y con el fin de respetar el límite de páginas establecido, no se reproducirá aquí nuevamente.

Rúbrica de corrección del examen

La tabla de abajo detalla los criterios de evaluación y la puntuación asignada a cada uno de los ejercicios, con el objetivo de asegurar una calificación coherente, transparente y objetiva.

Pregunta	**Criterio**	**Puntaje**
Tabulación cruzada	Ejercicio completo	2’5 pts
a)	Segmentos correctos	1 pts
	Gráfico	0’5pts
b)	Respuesta completa y argumentada	1pt
A. Correspondencias	Ejercicio completo	2’5 pts
a)	Asociación correcta	0’5pts
b)	Argumento válido	0’5pts
c)	Producto correcto	0’5pts

d)	Producto correcto	0’25pts
e)	Gráfico	0’25pts
f)	Argumento válido	0’5pts
WARD	Ejercicio completo	2’5pts
Apartados	Perfiles identificados	1’5pts
	Dendrograma	1pt
MDS Preferencias	Ejercicio completo	2’5pts
Apartados	Mapa perceptual	1pt
	Explicación mapa	1’5pts

Tabla 1. Rúbrica de la corrección del examen

Como se puede observar, cada alumno tendrá una nota sobre 10 en el examen, que constituye, como se especificó en el capítulo 3, el 70% de la nota final. Siendo necesario superar el 5 en el propio examen para aprobar la asignatura.

El 30% restante de la nota final corresponde a la evaluación continua, la cual se compone de ejercicios entregables explicados anteriormente. Cada uno de estos ejercicios tiene un peso específico de 3% de la nota final, ya que se considera un ejercicio por cada técnica explicada, tal como se detalla a continuación.

Ejercicio entregable	**Criterio**	**Puntaje**
Tabulación cruzada	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
LDA	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
QDA	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
A. Factorial	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
A. Correspondencias	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
K-means	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
WARD	Jupyter enviado	0’2 pts

	Formulario complementado	0,1pt
Semejanzas	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
Preferencias	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt
CHAID	Jupyter enviado	0’2 pts
	Formulario complementado	0,1pt

Tabla 2. Rúbrica de los ejercicios entregables

En esta puntuación, al ser parte de la evaluación continua, no hay mínimo y dependerá del docente considerar si es suficiente para que el alumno obtenga la nota completa de cada apartado el hecho de entregarlo o debe haber logrado obtener la solución correcta de cada ejercicio.

Todo este material didáctico está en la plataforma didáctica que explico en el siguiente capítulo.

Capítulo 6.
Plataforma educativa

Con el objetivo de facilitar el aprendizaje y optimizar la comunicación entre el alumnado y el docente se ha creado para este Trabajo de Fin de Grado una plataforma educativa digital. Es una herramienta pensada para ser un espacio centralizado para la entrega de actividades, el acceso a materiales complementarios, el seguimiento del progreso académico y la resolución de dudas. De esta forma tienen a su disposición un sistema donde pueden trabajar sin depender de la configuración de sistemas y software de sus respectivos equipos.

6.1. Estructura y componentes

Una plataforma educativa es un entorno digital que facilita la enseñanza y el aprendizaje mediante la gestión de contenidos, la interacción entre usuarios y el seguimiento del progreso académico. En este apartado se explicará de forma concisa la estructura de la misma. Para mantener limitada la extensión de la memoria, no vamos a explicar el sistema en el que esta soportada la arquitectura en este capítulo, la información al respecto está en el Anexo IV.

Dicho esto, a nivel funcional, la plataforma se estructura mediante una arquitectura modular basada en microservicios. Está soportada por un servidor Linux, basado en la última versión *lts* de Ubuntu, y hace uso de la plataforma Docker para desplegar un conjunto de contenedores.

Entre los componentes principales destacan dos pilares fundamentales:

- *Nextcloud*, gestor documental y de usuarios. En él se ha preparado un entorno para almacenar la información y las herramientas de evaluación de los usuarios.
- Jupyter*hub*, herramienta que habilita un entorno de desarrollo donde realizar los ejercicios propuestos.

Para poder acceder a estos dos pilares hacemos uso de una página web desplegada mediante el uso de *flask* en otro contendor, quedando nuestra arquitectura con el siguiente esquema:

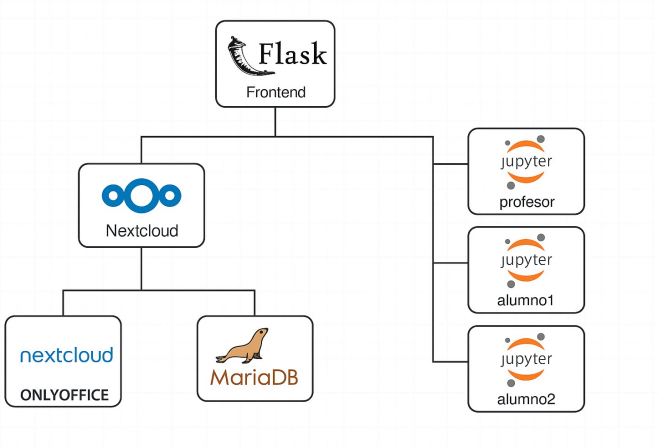


Ilustración 5: Arquitectura de la plataforma

Portal de inicio

Para empezar a trabajar con la plataforma lo primero que hacemos es acceder a la página de inicio:

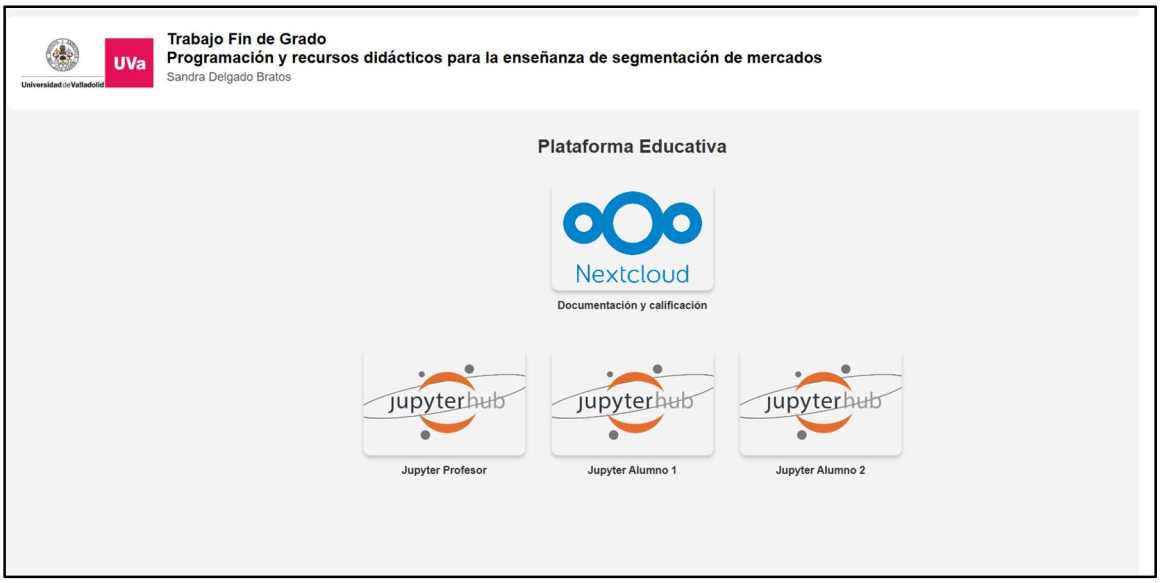


Ilustración 6: Portal de inicio de la plataforma

Como se puede ver en la imagen, desde el portal tenemos acceso a todos los sistemas necesarios para gestionar la información. Se ha decidido establecer tres entornos de desarrollo, profesor, alumno 1 y alumno 2; para la demostración de uso.

Comenzaremos revisando el gestor documental Nextcloud.

6.2. Nextcloud

Nextcloud es una solución de software libre que permite desplegar una nube privada de almacenamiento y colaboración en línea, similar a servicios comerciales como Google Drive o Dropbox, pero con un enfoque centrado en la privacidad y el control total por parte del usuario o la organización. Este sistema permite alojar archivos, compartir documentos, gestionar calendarios, contactos y realizar videollamadas, así como gestionar los permisos de accesos para cada alumno o grupo, todo dentro de una infraestructura propia.

Una de las principales ventajas es su arquitectura modular, que permite extender sus funcionalidades mediante aplicaciones, adaptándose así a distintas necesidades educativas, empresariales o personales. Además, al ser de código abierto, Nextcloud puede auditarse, modificarse y desplegarse en servidores propios, garantizando un entorno más seguro y conforme con legislaciones de protección de datos como el RGPD.

Para acceder a la plataforma procederemos a hacer clic encima del icono de Nextcloud. Tanto la primera vez en que accedemos como en cada ocasión en que no tenemos abierta ninguna sesión en el equipo, lo que vemos es la siguiente pantalla:

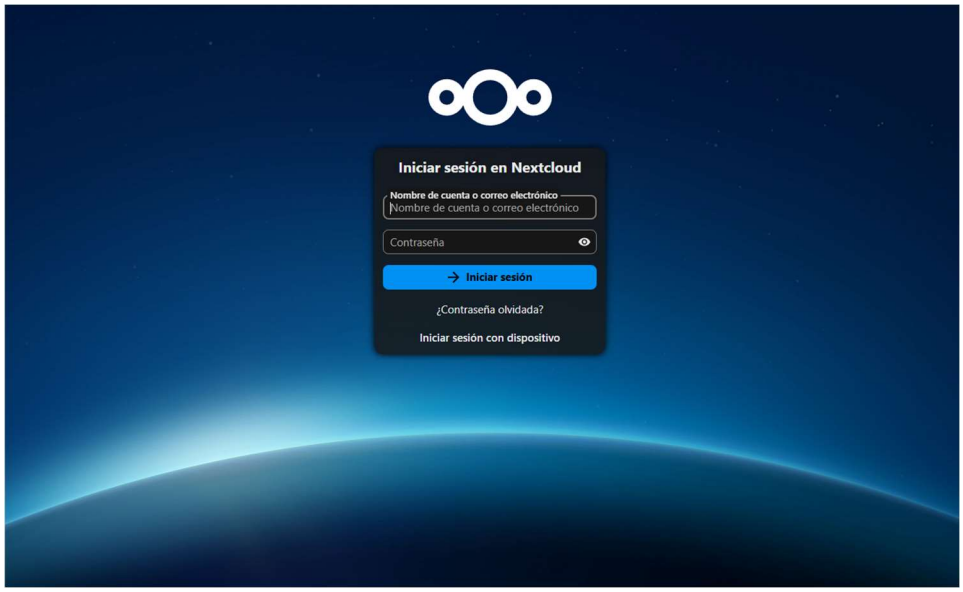


Ilustración 7: pantalla de autenticación de Nextcloud

Una vez introducidas nuestras credenciales de acceso, en el caso de ser administradores, accederemos a una pantalla en la que poder explorar los ficheros:

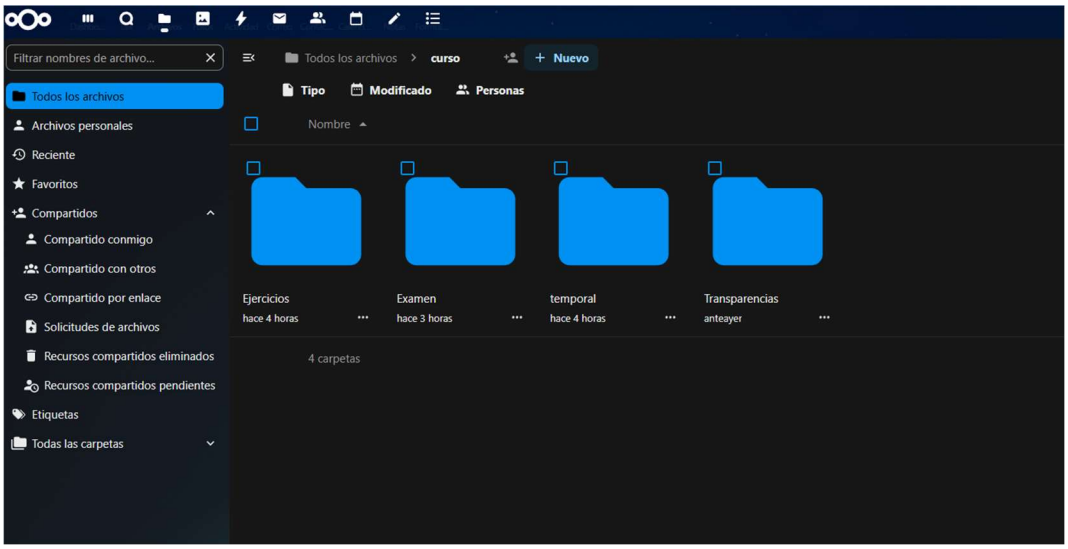


Ilustración 8: carpeta del curso abierta desde Nextcloud

Dentro del gestor, según nuestro nivel de acceso, podemos ir observando que tenemos diferentes funcionalidades.

La más importante es la funcionalidad de gestor de información. En la imagen anterior se puede apreciar como la información está organizada dentro de la plataforma y en la imagen posterior se ve el contenido de una de una de las carpetas.

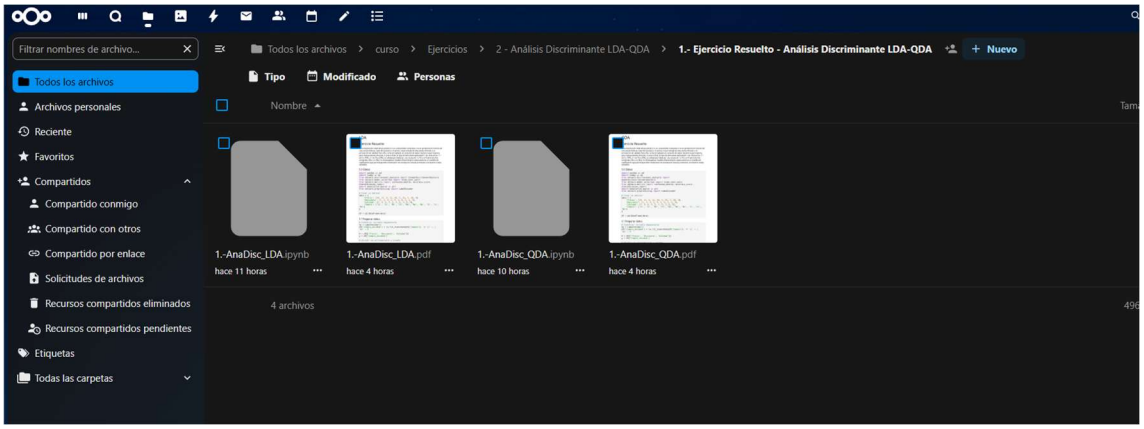


Ilustración 9: Vista de la carpeta de problemas resueltos para el análisis discriminante

Gracias a la capacidad de gestión de usuarios, tenemos organizadas las capas de visualización de la información de forma que nos permite discriminar tanto a quién como cuándo le es permitido acceder a la misma.

En este momento, se han creado cuatro usuarios: un profesor, que es el administrador, dos alumnos, alumno1 y alumno2, y un *observadordocente,* creado específicamente para poder visualizar el curso completo en el sistema. Podemos verlo en la siguiente imagen:

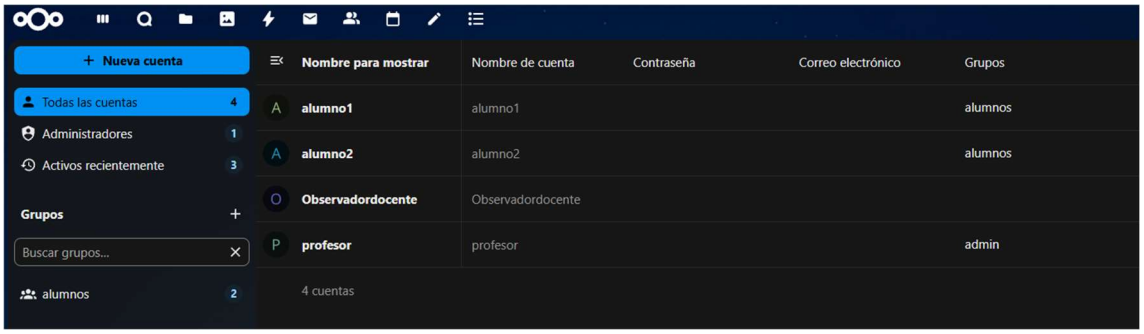


Ilustración 10: Ventana de administración de la gestión de usuarios

La tercera función es la de realización de evaluaciones mediante el uso de formularios. Estos formularios están diseñados para utilizarse como mecanismo de evaluación sobre cada una de las técnicas y pueden ser enviados a cualquier usuario de la plataforma. Además, es posible recuperar y analizar las respuestas de los mismos, por lo que agiliza el trabajo del docente.

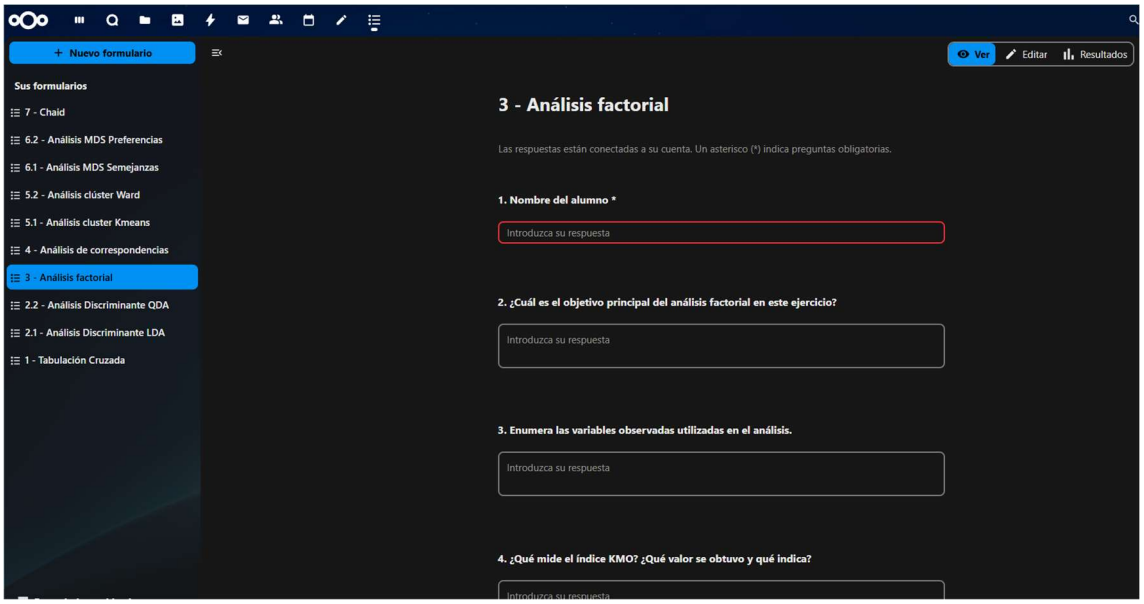


Ilustración 11: Gestor de formularios con uno de ellos abierto

Por último, el sistema que permite editar los ficheros dentro de la plataforma Nextcloud sin tener que descargarlos es el *onlyoffice* y está interconectado con Nextcloud. Esta integración permite a los usuarios mantener la centralización de los archivos y facilita el trabajo docente de forma que los alumnos pueden consultar los materiales o rellenar formularios directamente desde la propia plataforma y también los profesores pueden editar las guías o subir correcciones sin abandonar el entorno de trabajo.

6.3. Jupyterlab

JupyterLab es un entorno interactivo de desarrollo diseñado principalmente para trabajar con datos, código y documentos científicos. Forma parte del proyecto Jupyter y representa una evolución del clásico Jupyter Notebook, ofreciendo una interfaz más flexible, modular y potente. A través de JupyterLab, los usuarios pueden combinar celdas de código ejecutable (en lenguajes como Python, R o Julia), texto enriquecido en formato Markdown, gráficos dinámicos y salidas interactivas, todo en un mismo espacio de trabajo.

Una de las características más destacadas de JupyterLab es su arquitectura basada en extensiones, lo que permite personalizar la experiencia del usuario y añadir funcionalidades adicionales según las necesidades del proyecto. Su integración con librerías de visualización y procesamiento de datos lo convierte en una herramienta muy utilizada en ciencia de datos, investigación académica y educación superior.

Al ser una plataforma de código abierto, JupyterLab puede instalarse y ejecutarse tanto en entornos locales como en servidores remotos, lo que facilita su adopción en infraestructuras educativas o científicas que requieren control total sobre los entornos de ejecución.

Este es el principal entorno de trabajo para los alumnos y permite desarrollar ejercicios interactivos de programación, análisis de datos y visualización, todo desde el navegador y sin necesidad de configurar nada en su equipo local. El entorno está proporcionado directamente desde la plataforma. Su integración en la plataforma hace posible que los alumnos trabajen sobre los materiales que encuentran en Nextcloud y entreguen sus soluciones de forma centralizada y controlada.

Dispone de un control de acceso, al igual que Nextcloud, para poder segurizar el entorno y que solo el usuario para el que está configurado pueda acceder a la aplicación. Podemos ver a continuación la ventana de autentificación:

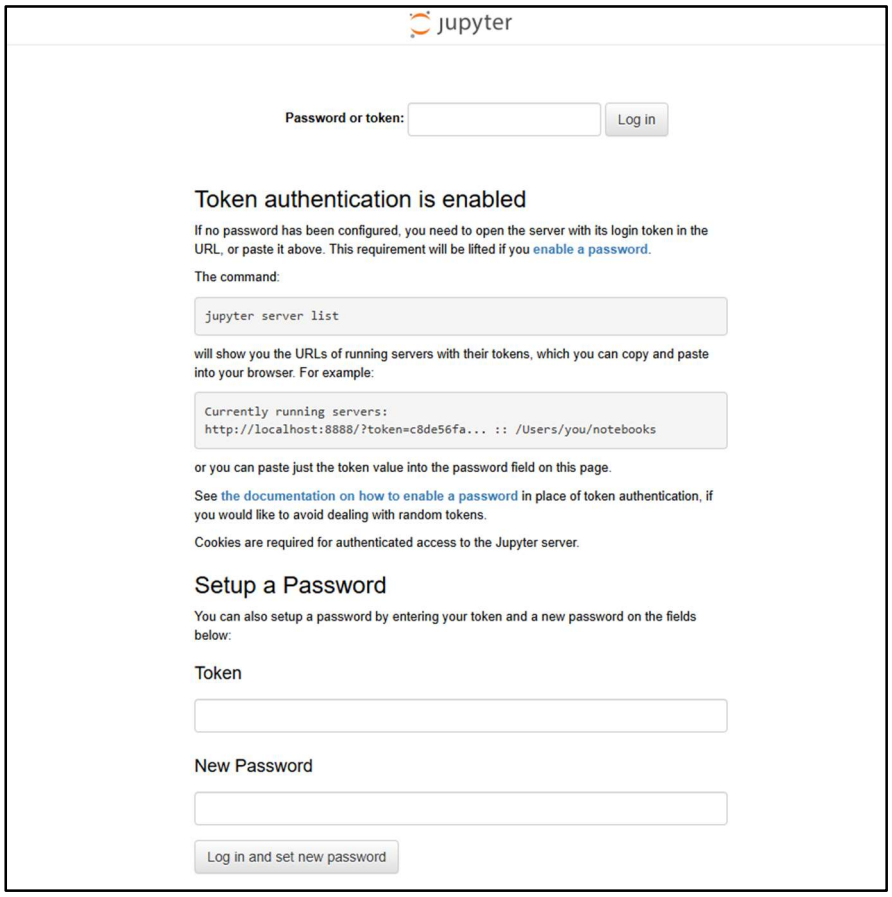


Ilustración 12: Ventana de autentificación de Jupyterlab

Una vez introducimos las credenciales de acceso, accedemos a la interfaz de programación propiamente dicha:

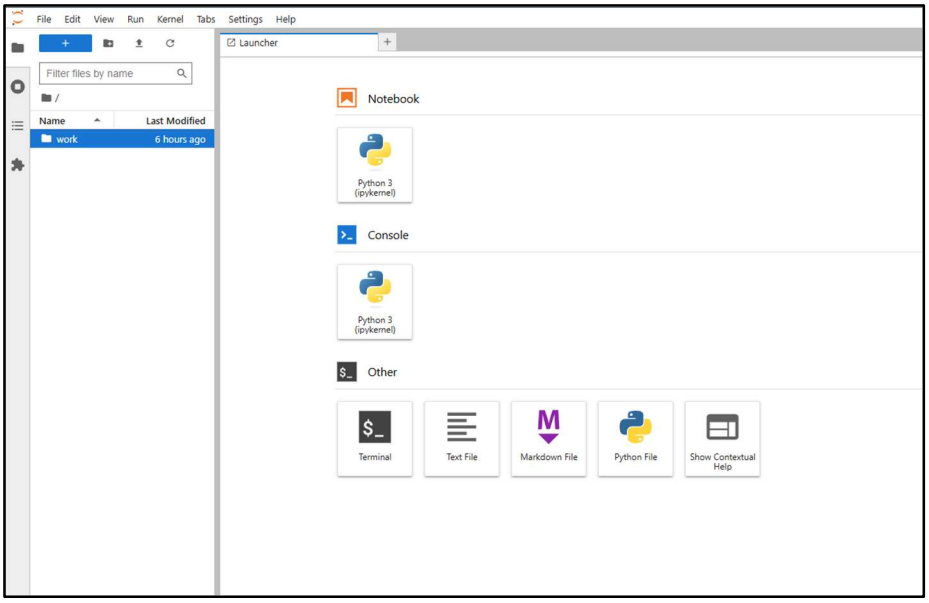


Ilustración 13: Ventana de trabajo de JupyterLab

En esta ventana ya se puede acceder a las distintas funcionalidades que permite este entorno y donde principalmente podremos cargar y ejecutar los Jupyter notebooks.

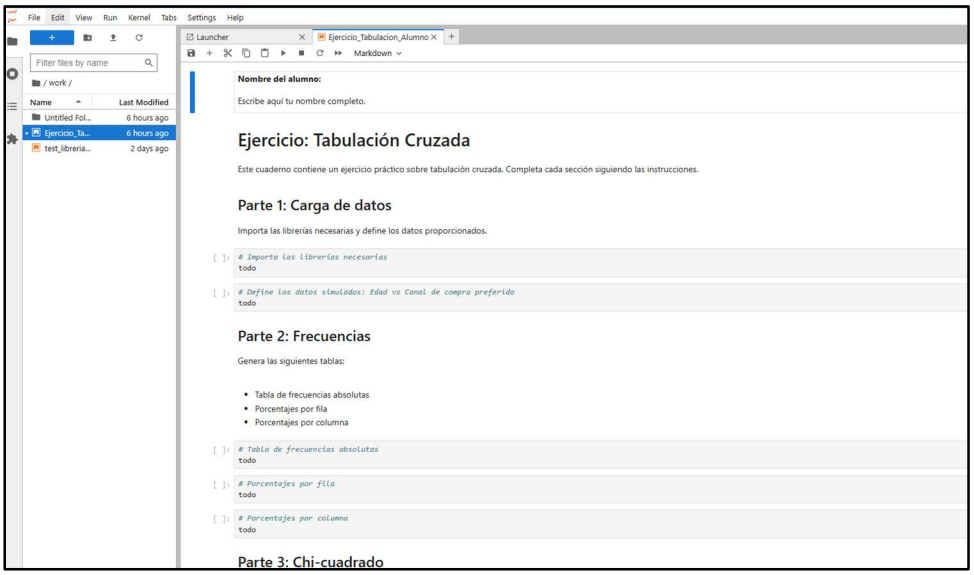


Ilustración 14: Cuaderno de tabulación cruzada listo para ser rellenado por un alumno

Toda la información técnica de esta plataforma que no ha sido especificada en este capítulo está reunida en el Anexo IV.

El servidor es accesible a través de la IP pública http://51.83.6.61

Credenciales de acceso

A continuación, se detallan los puertos de acceso, credenciales y descripción de cada uno:

Servicio	Puerto externo	Usuario	Contraseña / Token
frontend	5000	—	—
db (MariaDB)	—	nextcloud	••••••••
nextcloud	8080	profesor	profesor
nextcloud	8080	alumno1	alumno1alumno1
nextcloud	8080	alumno2	alumno2alumno2
Nextcloud	8080	observadordocente	observadordocente
jupyteralumno1	8889	alumno1	alumno1
jupyteralumno2	8890	alumno2	alumno2
jupyterprofesor	8888	profesor	profesor
onlyoffice	9980	—	supersecret

Capítulo 7.
Conclusiones

El objetivo principal de este Trabajo de Fin de Grado fue, como se especificó en el apartado 1.2, la creación de un documento que sirva como material didáctico integral y actualizado para la enseñanza de la segmentación de mercados en la universidad incorporando herramientas digitales y tecnológicas.

Para ello se han desarrollado los contenidos teóricos de la segmentación de mercados partiendo de conocimientos generales hasta llegar al concepto y el uso de la misma. También se han explicado las técnicas estadísticas con las que se puede llevar a cabo la segmentación y se han organizado de manera que el estudio parta de las más sencillas a las más complejas.

Se ha planteado un proyecto formativo completo y se han creado los recursos para ello, siendo estos recursos: presentaciones para el aula, problemas resueltos para guiar la explicación con un ejemplo, problemas propuestos para que el alumno pueda practicar, pruebas evaluables para llevar a cabo la evaluación continua y un examen final con el que se evaluarán los conocimientos adquiridos.

Además, uno de los principales aportes de este trabajo es la propia creación de una plataforma educativa digital que centraliza todo el material didáctico desarrollado, permitiendo su aplicación inmediata en contextos reales de enseñanza. La plataforma se presenta como una propuesta coherente y sostenible. Su diseño versátil y adaptable la convierte en un modelo replicable en otros niveles o áreas educativas, ofreciendo así una contribución significativa tanto a la práctica docente como a futuras investigaciones o proyectos en el ámbito de la innovación educativa.

Por tanto, puede afirmarse que los objetivos planteados al inicio del trabajo se han cumplido en su totalidad, especialmente en lo referente a la creación de un recurso formativo completo y su implementación digital.

7.1. Posibles mejoras

Una posible continuación de este trabajo podría ser la instalación de nuevas funcionalidades en la plataforma educativa o ampliar las técnicas estadísticas en el caso de que hubiera la posibilidad de ampliar el tiempo de enseñanza.

También sería interesante plantear la gamificación y otras metodologías activas que pueden integrarse dentro de la plataforma para enriquecer más la experiencia de aprendizaje.

7.2. Limitaciones

Cabe destacar que durante el desarrollo del trabajo se identificaron algunas limitaciones como la duración esperada de la enseñanza de este material, tanto teórica como práctica, que se estableció de uno a tres meses. Siendo así, no era posible ahondar en todas las técnicas de segmentación de mercados que se hubiera querido ni con la profundidad deseada. Por ello, se ha escogido una selección de las mismas que crean una base de conocimientos completa para el alumnado y su explicación se ha diseñado con un enfoque más práctico.

7.3. Dificultades superadas

A lo largo del desarrollo del trabajo surgieron retos como la selección de las técnicas estadísticas más adecuadas al tiempo disponible, la organización coherente del contenido teórico y la creación de una plataforma digital funcional con herramientas que satisficieran los requisitos de eficiencia, escalabilidad, gratuidad de uso y extensibilidad.

Estas dificultades fueron superadas mediante una investigación exhaustiva de literatura académica, así como a través de la exploración y evaluación de soluciones digitales disponibles en línea que satisfacían los criterios funcionales establecidos. El proceso también implicó un trabajo continuo de revisión y ajuste de la secuencia didáctica, hasta lograr una estructura lógica que facilitara el aprendizaje progresivo por parte del alumnado.

7.4. Reflexiones finales

Este trabajo ha supuesto una oportunidad para integrar y aplicar conocimientos adquiridos en asignaturas como *Análisis de Datos, Análisis Multivariante, Análisis de Datos Categóricos, Fundamentos de Sistemas Operativos,* así como aquellas orientadas a la programación en distintos lenguajes. Todo ello ha evidenciado la interconexión entre teoría y práctica.

Durante el desarrollo del trabajo también he adquirido nuevos aprendizajes, como la creación de herramientas tecnológicas aplicadas a la docencia, la planificación didáctica y la capacidad de adaptar contenidos técnicos a formatos pedagógicos accesibles.

Referencias

A, G., & Syam, R. (25 de Enero de 2024). *Using DBSCAN to identify customer segments with high churn risk on Amazon consumer behavior data.* Obtenido de TechRxiv: 377718167_Using_DBSCAN_to_Identify_Customer_Segments_with_High_Churn_Risk_on_Amazon_Consumer_Behavior_Data

De la Fuente Fernández, S. (2011). *Análisis de correspondencias simples y múltiples.* Obtenido de www.fuenterrebollo.com: https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/REDUCIR-DIMENSION/CORRESPONDENCIAS/correspondencias.pdf

Escobar, A. (1998). Whose knowledge, whose nature? Biodiversity, conservation, and the political ecology of social movements. *Journal of Political Ecology*, 53 – 76. Obtenido de https://sociocav.usal.es/blog/modesto-escobar/files/2014/06/Escobar1998a.pdf

Flury, B. (1997). *A First Course in Multivariate Statistics.* New York: Springer.

Grande Esteban, I. &. (2017). *Fundamentos y técnicas de investigación comercial.* Madrid: ESIC.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Kuo, Y.-F., & Yang, C.-H. (2022). Data-driven market segmentation in hospitality using unsupervised learning techniques. *Tourism Management Perspectives*, 100991.

Lieder, I., Segal, M., Avidan, E., Cohen, A., & Hope, T. (2019). Learning a faceted customer segmentation for discovering new business opportunities at Intel. *Proceedings of IEEE BigData 2019* (págs. 93–101). Los Ángeles: IEEE.

Lozano Ruiz, I. (2025). *Optimización de campañas de marketing en un centro de comercio mediante segmentación de clientes.* Bogotá: UNAD.

Ortega Martínez, E. (1992). *Manual de investigación comercial.* Madrid: Pirámide.

Peña, D. (2002). *Análisis de datos multivariantes.* Madrid: McGraw Hill.

Roche, I. C. (1991). *Fundamentos de Marketing.* Barcelona: Ariel.

Rodríguez Palacios, J. E., & López Rivera, J. R. (2025). *Preferencias y hábitos de consumo del mercado de café tostado en el Distrito Central, Francisco Morazán.* Tegucigalpa: UNITEC. Obtenido de https://repositorio.unitec.edu/items/d890a8a2-71c2-436d-866d-b0d87f6dd736

Rothman, A. J., Bartels, R. D., Wlaschin, J. M., & S. P. (2006). Use of the Risk Perception Attitude Framework in HIV Prevention in the United States. *American Journal of Public Health*, 2224–2229.

Santesmases Mestre, M. (1996). *Marketing: conceptos y estrategias.* Madrid: Pirámide.

Seber, G. A. (1984). *Multivariate Observations.* New York: Wiley.

Smith, W. R. (1956). Wendell R. Smith,Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, 3-8.

Srivastava, M. (2002). *Methods of Multivariate Statistics.* New York: Wiley.

Zhang, T. (Enero de 2024). Research on demographic segmentation and marketing strategies for young people—Based on the case study of Coca-Cola. *Advances in Economics, Management and Political Sciences*, 93–101.

Obtenido de https://www.researchgate.net/publication/377180005_Research_on_Demographic_Segmentation_and_Marketing_Strategies_for_Young_PeopleBased_on_the_Case_Study_of_Coca-Cola

Índice de figuras

Índice de tablas

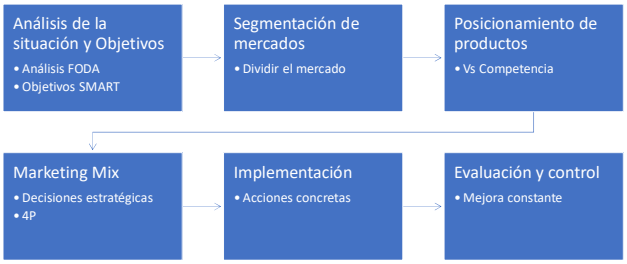
Anexo I.
Presentaciones para el aula

Investigación de mercado

- Obtención sistemática de información para asistir a la dirección en la toma de decisiones comerciales de un problema concreto.

Marketing Estratégico

- Enfoque a largo plazo en la planificación y ejecución de actividades de marketing que busca crear y mantener una ventaja competitiva sostenible.



Definición: Segmentación de mercados

- La segmentación de mercados es una herramienta del marketing estratégico que utiliza el análisis estadístico para dividir un mercado amplio y heterogéneo en subgrupos más pequeños y homogéneos.

- Su principal objetivo es maximizar la eficiencia y la efectividad de las actividades de marketing al adaptar productos, precios, comunicación y canales de distribución a las características únicas de cada segmento.

Principios clave de la segmentación

- Diversidad del consumidor
- Optimización de recursos
- Mejora de la satisfacción del cliente
- Diferenciación competitiva
- Facilitación del desarrollo de productos
- Adaptabilidad y respuesta al cambio

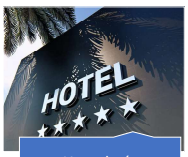
Aplicación real

Coca-Cola

Amazon

Intel

Salud pública en Malawi



Hostelería

Principios para una segmentación de mercado

Requisitos

UVa

Para la segmentación	Para el segmento
• Valoraciones diferenciadas de los consumidores • Vinculación de las diferencias percibidas a grupos de consumidores • Posibilidad de implementar marketing diferenciado • Potencial de ventas suficiente • Estabilidad en el tiempo	• Medible • Rentable • Accesible • Factible • Fácilmente identificable • Heterogéneo • Defendible

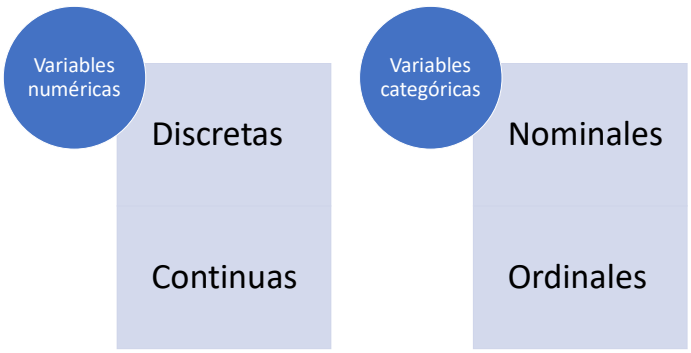
Criterios de segmentación

UVa

Generales Objetivos	• Segmentación Demográfica y Socioeconómica • Segmentación Geográfica
Generales Subjetivos	• Segmentación Psicográfica • Segmentación Conductual
Específicos Objetivos	• Segmentación por Uso
Específicos Subjetivos	• Segmentación por Beneficio

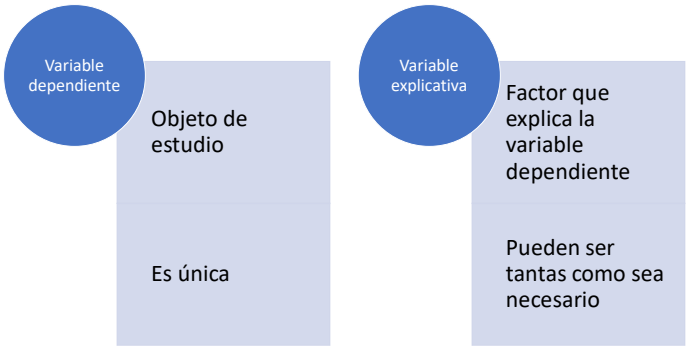
Variables

UVa



Variables

UVa



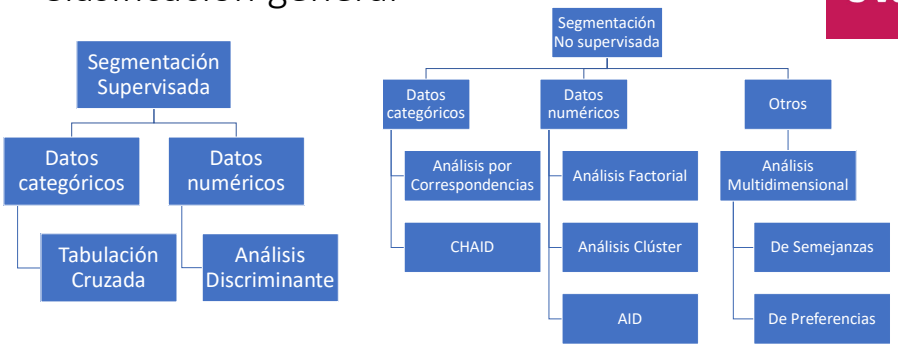
Proceso de segmentación de mercados

UVa

UVa

Técnicas estadísticas de segmentación

Clasificación general



Segmentación Supervisada

Datos categóricos

Tabulación cruzada

La cafetería Gran Celebración ha decidido estudiar las ventas de sus bebidas en función de dos variables: sexo [hombre, mujer] y la bebida [café, té, zumo o frappé]. Los resultados obtenidos fueron:

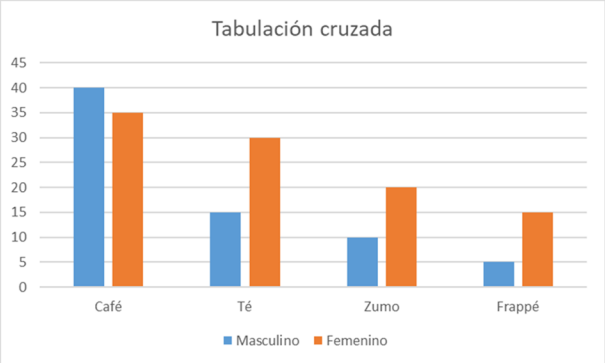
Bebida / Sexo	Café	Té	Zumo	Frappé	Total
Masculino	40	15	10	5	**70**
Femenino	35	30	20	15	**100**
Total	**75**	**45**	**30**	**20**	**170**

Frecuencias porcentuales

Bebida / Sexo	Café	Té	Zumo	Frappé	Total
Masculino	23,52941	8,823529	5,882353	2,941176	**41,17647**
Femenino	20,58824	17,64706	11,76471	8,823529	**58,82353**
Total	**44,11765**	**26,47059**	**17,64706**	**11,76471**	**100**

- Acuden más mujeres que hombres.
- La bebida más tomada es el café.
- Más de la mitad de los hombres toman café, mientras que las mujeres no tienen una preferencia.

Gráficamente



Test Chi-Cuadrado de independencia

- Tamaño de muestra suficientemente grande (frecuencias esperadas en cada celda de 5 o más)
- Observaciones independientes
- H0: Tipo de bebida y sexo del cliente no están relacionados.
- H1: Tipo de bebida y sexo del cliente sí están relacionados.

Bebida / Sexo	Café	Té	Zumo	Frappé
Masculino	40	15	10	5
Femenino	35	30	20	15

Frecuencias esperadas

UVa

$$f_i(\mathrm{i}, j) = \frac{R_i \times C_j}{N}$$

Bebida / Sexo	Café	Té	Zumo	Frappé	Total
Masculino	30,88235	18,52941	12,35294	8,235294	**70**
Femenino	44,11765	26,47059	17,64706	11,76471	**100**
Total	**75**	**45**	**30**	**20**	**170**

Estadístico Chi-cuadrado

UVa

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{Eij}$$

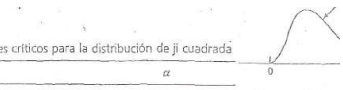
χ^2 = 8,641666667

Grados de libertad

$$gl = (i-1) \times (j-1) = 3$$

Valor Crítico

Tabla A.7 Valores críticos para la distribución de ji cuadrada



ν	.995	.99	.975	.95	.90	.10	.05	.025
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735

P-valor

UVa

Es un valor entre 0 y 1 que indica la probabilidad de que la hipótesis nula es verdadera.

- Si el p-valor es menor que 0.05, rechazamos la hipótesis nula.
- Si el p-valor es mayor que 0.05, decimos que no se puede rechazar la hipótesis nula, pues no hay diferencias significativas.

Ejercicio: TCruzada1.ipynb

A partir de aquí todos los ejercicios
se ejecutarán en jupyter notebook

Segmentación Supervisada

Datos numéricos

LDA

Análisis Discriminante Lineal

- Función discriminante para una clase k :

$$\delta k(x)=x^T*\Sigma-1\mu k-21\mu k^T\neg*\Sigma-1\mu k+\log(\pi k)$$

- Suposiciones:
  - Normalidad Multivariante
  - Igualdad de Covarianzas
  - Linealidad

Ejercicio: LDA1.ipynb

QDA

Análisis Discriminante Cuadrático

- Función discriminante para una clase k :

$$\delta k(x) = -21 \log|\Sigma k| - 21 (x-\mu k)T\Sigma k-1(x-\mu k) + \log(\pi k)$$

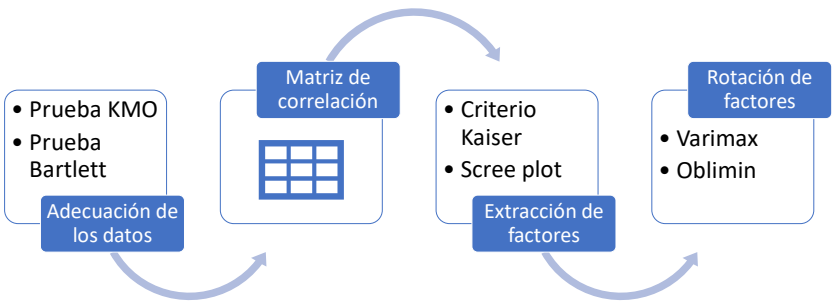
- Suposiciones:
  - Normalidad Multivariante
- Riesgo de sobreajuste:
  - Nº datos = 10 x p_variables x k_clases

Ejercicio: QDA1.ipynb

Segmentación No Supervisada

Análisis Factorial

Objetivo: Reducir la dimensionalidad.



Análisis Factorial

UVa

Una empresa de telecomunicaciones busca segmentar a sus usuarios en base a comportamientos observados y no sus preferencias personales.

Usuario	Minutos al mes	Número de llamadas	Datos móviles (GB)
U1	800	200	3.5
U2	750	180	4.0
U3	300	50	12.0
U4	280	45	11.0
U5	700	190	2.0

Ejercicio: Factorial1.ipynb

CA

UVa

Análisis por Correspondencias

Es una ampliación de la tabulación cruzada con técnicas del análisis factorial.

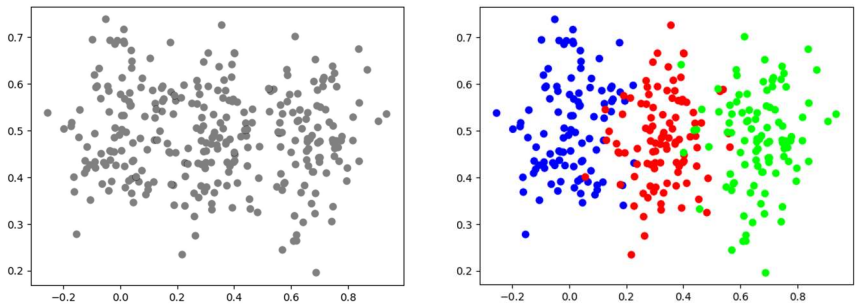
Objetivos:

- Reducir la dimensionalidad
- Representar gráficamente las relaciones entre categorías de las variables
- Identificar asociaciones y patrones

Ejercicio: Corr1.ipynb

Análisis Clúster

UVa



Tipos de clustering

UVa

Clustering Particional	Clustering Jerárquico	Basados en Densidad	Basados en Distribución
• K-Means	• Linkage Method: - Ward	• DBSCAN	• GMM

K-Means

UVa

- Divide los datos en k clústeres.
- Minimiza la suma de las distancias cuadradas entre los puntos y el centroide del clúster.

- Es rápido y fácil de implementar.
- Requiere especificar el número de clústeres k.
- Puede converger a un óptimo local.

Ejercicio: Kmeans1.ipynb

Ward

UVa

- Es un método de enlace.
- Produce un dendrograma que muestra la fusión o división de clústeres.

- No requiere el número de clústeres de antemano.

Ejercicio: Ward1.ipynb

MDS

UVa

Análisis Multidimensional

- Datos subjetivos
- Buscamos representar visualmente las percepciones de los encuestados

MDS métrico	MDS no métrico
• Disimilitudes numéricas • Conservar distancias reales	• Orden o ranking de disimilitudes • No se conservan proporciones exactas

Semejanza

UVa

- Los datos son pares de productos, marcas o atributos.

Hicimos una encuesta a consumidores, y obtuvimos la siguiente matriz de disimilitudes:

Escala 1–9, donde
1 = muy parecidos
9 = muy distintos

Ejercicio: Semejanza1.ipynb

	A	B	C	D
A	0	3	7	6
B	3	0	5	4
C	7	5	0	2
D	6	4	2	0

Preferencia

UVa

- Los datos son preferencias individuales respecto a un conjunto de productos, marcas o atributos.

3 consumidores califican 4 productos (A, B, C, D) :

Escala de 1 a 10
1 = menor preferencia
10 = mayor preferencia

Ejercicio: Preferencia1.ipynb

Producto	Persona 1	Persona 2	Persona 3
A	8	7	6
B	6	5	4
C	4	8	7
D	2	3	5

AID

Automatic Interaction Detection

- Crea un árbol de decisión
- Parte de una variable dependiente numérica
- Utiliza ANOVA para hacer las divisiones
  - Divisiones binarias
  - Maximiza la diferencia en promedios:

- Previsión de gasto de un cliente.

CHAID

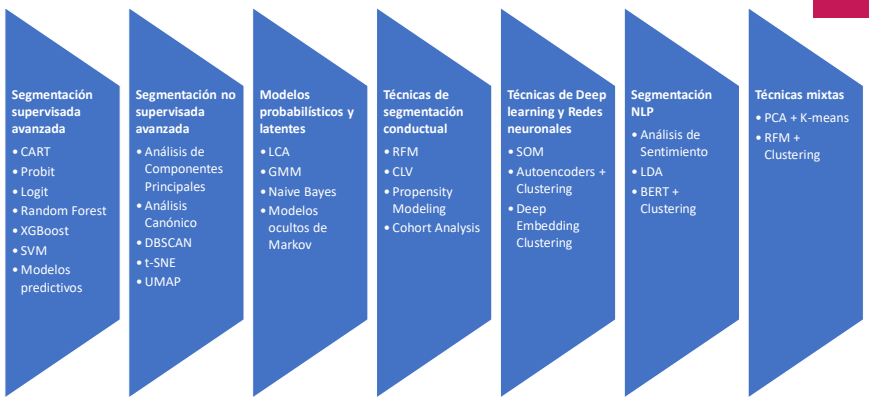
Chi-Squared Automatic Interaction Detection

- Crea un árbol de decisión
- Parte de una variable dependiente categórica
- Utiliza test Chi-cuadrado para hacer las divisiones
  - Permite divisiones múltiples
  - Segmentos significativamente distintos:

- Segmentación de mercado.
- Análisis de comportamiento.

Ejercicio: CHAID.ipynb

Otros modelos



Anexo II.
Problemas resueltos

Ejercicio de tabulación cruzada y chi cuadrado

La cafetería Gran Celebración ha decidido estudiar las ventas de sus bebidas en función de dos variables: sexo [hombre, mujer] y la bebida [café, té, zumo o frappé]. Los resultados obtenidos fueron:

Bebida / Sexo	Café	Té	Zumo	Frappé	**Total**
Masculino	40	15	10	5	**70**
Femenino	35	30	20	15	**100**
Total	**75**	**45**	**30**	**20**	**170**

a) Analizar los posibles segmentos aplicando la técnica de tabulación cruzada, señalando los segmentos que considere más importantes. Incluye al menos un gráfico.
b) Comprobar si los posibles segmentos están o no relacionados.
c) ¿Qué acción sugieres realizar tras este análisis para aumentar las ventas?

Nota: El valor de la chi cuadrado al 95% de confianza, con 1 grado de libertad es 3,84; con 2 grados de libertad es 5,99; y con 3 grados de libertad es 7,81. Considere trabajar con 2 cifras decimales.

Ejercicio resuelto

La cafetería Gran Celebración ha decidido estudiar las ventas de sus bebidas en función de dos variables: sexo [hombre, mujer] y la bebida [café, té, zumo o frappé].

1.0 Datos

```
# Importar librerías necesarias
import pandas as pd import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency

# Crear el DataFrame con los datos
data = {
'Sexo': ['Masculino'] * 4 + ['Femenino'] * 4,
'Bebida': ['Café', 'Té', 'Jugo', 'Frappé'] * 2,
'Frecuencia': [40, 15, 10, 5, 35, 30, 20, 15]
}

df = pd.DataFrame(data)

# Crear tabla de frecuencias absolutas
tabla_frecuencia = pd.pivot_table(df, values='Frecuencia', index='Sexo', columns='Bebida', aggfunc=sum, fill_value=0)
```

```
C:\Users\San\AppData\Local\Temp\ipykernel_5136\3483513523.py:17: FutureWarning: The provided callable <built-in function sum> is currently using DataFrameGroupBy.sum. In a future version of pandas, the provided callable will be used directly. To keep current behavior pass the string "sum" instead.
  tabla_frecuencia = pd.pivot_table(df, values='Frecuencia', index='Sexo', columns='Bebida', aggfunc=sum, fill_value=0)
```

1.1 Frecuencias porcentuales

```
# Tabla de frecuencias porcentuales (por fila) tabla_porcentajes_fil = tabla_frecuencia.div(tabla_frecuencia.sum(axis=1), axis=0) * 100

# Tabla de frecuencias porcentuales (por columna) tabla_porcentajes_col = tabla_frecuencia.div(tabla_frecuencia.sum(axis=0), axis=1) * 100

# Tabla de frecuencias porcentuales (global)
```

```
tabla_porcentajes_global = (tabla_frecuencia / tabla_frecuencia.va-
lues.sum()) * 100
```

1.2 Test Chi-Cuadrado

```
# Prueba chi-cuadrado

chi2, p, dof, expected = chi2_contingency(tabla_frecuencia)
```

1.3 Resultados

```
# Mostrar resultados
print("Tabla de frecuencias absolutas:\n", tabla_frecuencia, "\n") print("Tabla de porcentajes por fila (%):\n", ta-
bla_porcentajes_fil.round(1), "\n")
print("Tabla de porcentajes por columna (%):\n", tabla_porcentajes_col.round(1), "\n") print("Tabla de porcentajes
globales (%):\n", tabla_porcentajes_global.round(1), "\n")
print(f"Chi-cuadrado = {chi2:.2f}, p-valor = {p:.4f}, grados de libertad = {dof}", "\n")
print("Tabla de frecuencias esperadas bajo H0:\n", pd.DataFrame(expected, index=tabla_frecuencia.index, co-
lumns=tabla_frecuencia.columns).round(2))
```

Tabla de frecuencias absolutas:

Bebida	Café	Frappé	Jugo	Té	Sexo
Femenino	35	15	20	30	
Masculino	40	5	10	15	

Tabla de porcentajes por fila (%):

Bebida	Café	Frappé	Jugo	Té	Sexo
Femenino	35.0	15.0	20.0	30.0	
Masculino	57.1	7.1	14.3	21.4	

Tabla de porcentajes por columna (%):

Bebida	Café	Frappé	Jugo	Té	Sexo
Femenino	46.7	75.0	66.7	66.7	
Masculino	53.3	25.0	33.3	33.3	

Tabla de porcentajes globales (%):

Bebida	Café	Frappé	Jugo	Té	Sexo
Femenino	20.6	8.8	11.8	17.6	
Masculino	23.5	2.9	5.9	8.8	

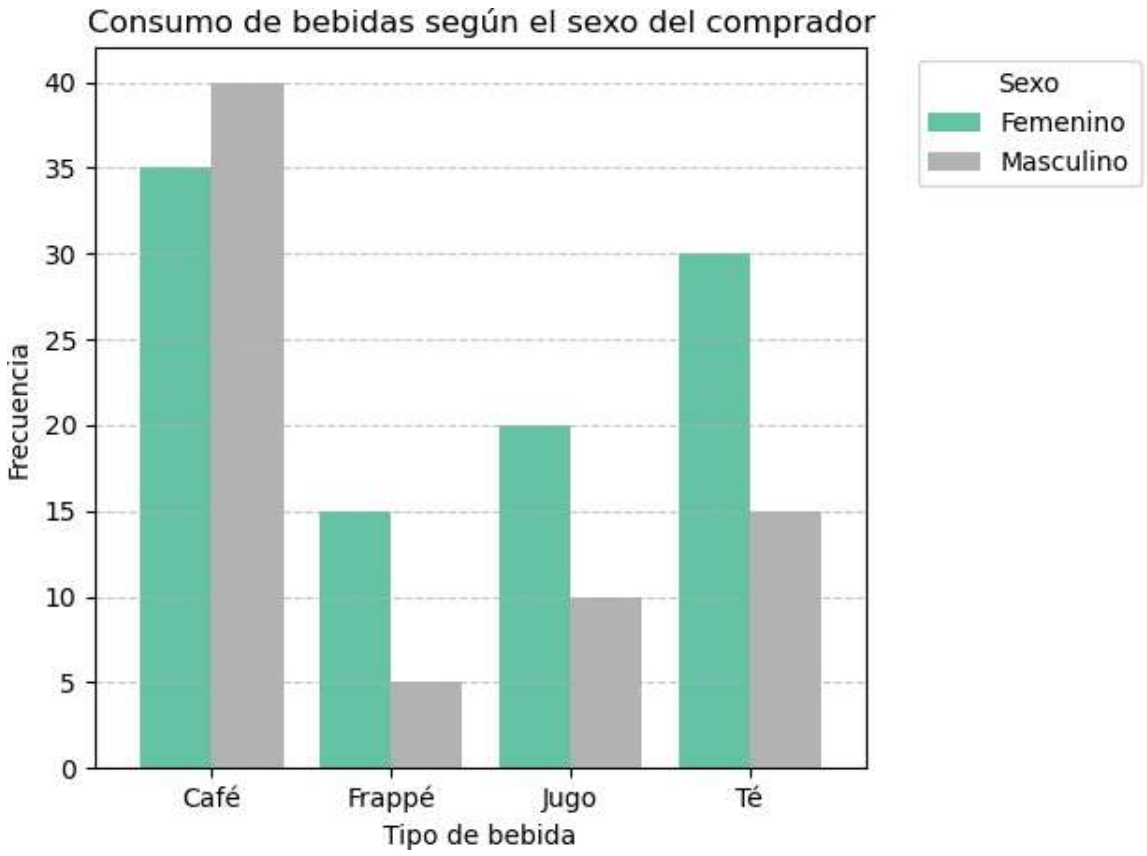
Chi-cuadrado = 8.64, p-valor = 0.0345, grados de libertad = 3

Sexo					
Femenino	44.12	11.76	17.65	26.47	Té
Masculino	30.88	8.24	12.35	18.53	

1.4

```
# Gráfico de barras no apiladas

plt.figure(figsize=(10, 6))
tabla_frecuencia.T.plot(kind='bar', colormap='Set2', width=0.8)
plt.title('Consumo de bebidas según el sexo del comprador')
plt.ylabel('Frecuencia')
plt.xlabel('Tipo de bebida') plt.xticks(rota-
tion=0)
plt.legend(title='Sexo', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



a) Analizar los posibles segmentos aplicando la técnica de tabulación cruzada, señalando los segmentos que considere más importantes. Incluye al menos un gráfico.

Podemos segmentar:

- Hombres + Café
- Mujeres + Resto

b) Comprobar si los posibles segmentos están o no relacionados.

p-valor = 0.0345 < 0.05 Rechazamos la H0: las variables son independientes. Por lo tanto, una influye en la otra y es recomendable segmentar.

c) ¿Qué acción sugieres realizar tras este análisis para aumentar las ventas?

- Crear una bebida de café que atraiga a las mujeres
- Mejorar la oferta de las otras bebidas para el público masculino, creando menús en las que se incluyan
- Hacer un nuevo estudio con el horario de venta de las bebidas
- Ampliar / Reducir / Cambiar la oferta de Frappé / Zumo Nota: depende del cliente objetivo, esta es una respuesta libre.

Ejercicio de Análisis Discriminante – LDA y QDA

Una empresa de *retail* desea predecir si un consumidor comprará o no un producto en función de tres características clave del producto: el precio, el porcentaje de descuento ofrecido y la percepción de calidad. Para ello, se ha recopilado un conjunto de datos históricos que registra, para cada producto ofrecido, el precio final, el tipo de descuento aplicado(0 = sin descuento, 1 = de 5 a 10%, 2
= de 15 a 25%), la calidad percibida (en una escala de 1 a 10) y si finalmente fue comprado (Sí) o no (No).

Se desea aplicar Análisis Discriminante para construir un modelo de clasificación que permita predecir la decisión de compra de futuros productos con base en estas variables.

Precio	10	15	8	12	20	5	25	7	30	9
Descuento	1	2	0	2	0	1	0	2	1	3
Calidad	7	8	6	9	5	4	6	7	8	5
Compra	Si	Si	No	Si	No	No	No	Si	Si	No

Aplica el análisis discriminante linear y cuadrático y compara resultados.

a) ¿Qué variables contribuyen más a la probabilidad de compra y cómo?
b) Haz un gráfico para visualizar la segmentación.
c) ¿El modelo categoriza bien los nuevos productos?
d) ¿Existe un umbral de precio a partir del cual la probabilidad de compra disminuye drásticamente?

Ejercicio Resuelto

Una empresa de retail desea predecir si un consumidor comprará o no un producto en función de tres características clave del producto: el precio, el porcentaje de descuento ofrecido y la percepción de calidad. Para ello, se ha recopilado un conjunto de datos históricos que registra, para cada producto ofrecido, el precio final, el tipo de descuento aplicado(0 = sin descuento, 1 = de 5 a 10%, 2 = de 15 a 25%), la calidad percibida (en una escala de 1 a 10) y si finalmente fue comprado (Sí) o no (No). Se desea aplicar Análisis Discriminante para construir un modelo de clasificación que permita predecir la decisión de compra de futuros productos con base en estas variables.

1.0 Datos

```
import pandas as pd im-
port numpy as np
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score, clas-
sification_report
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder

# Crear un dataset

data = {
'Precio': [10, 15, 8, 12, 20, 5, 25, 7, 30, 9],
'Descuento': [1, 2, 0, 2, 0, 1, 0, 2, 1, 3],
'Calidad': [7, 8, 6, 9, 5, 4, 6, 7, 8, 5],
'Compra': ['Sí', 'Sí', 'No', 'Sí', 'No', 'No', 'No', 'Sí', 'Sí', 'No']
}

df = pd.DataFrame(data)
```

1.1 Preparar datos

```
# Codificar variable dependiente

le = LabelEncoder()
df['Compra_encoded'] = le.fit_transform(df['Compra']) # 'Sí' → 1, 'No' → 0

X = df[['Precio', 'Descuento', 'Calidad']]
y = df['Compra_encoded']

# Dividir en entrenamiento y prueba
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)
```

1.2 Entrenar modelo

```
# Entrenar modelo LDA

lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)

LinearDiscriminantAnalysis()
```

1.3 Predicciones

```
# Predicciones

y_pred = lda.predict(X_test)
```

1.4 Evaluación

```
# Evaluación

print("Matriz de Confusión:") print(confusion_matrix(y_test, y_pred))

print("\nReporte de Clasificación:") print(classification_report(y_test, y_pred))

print(f"Score: {accuracy_score(y_test, y_pred):.2f}")
```

```
Matriz de Confusión:

[[1 0]

 [0 2]]

Reporte de Clasificación:
```

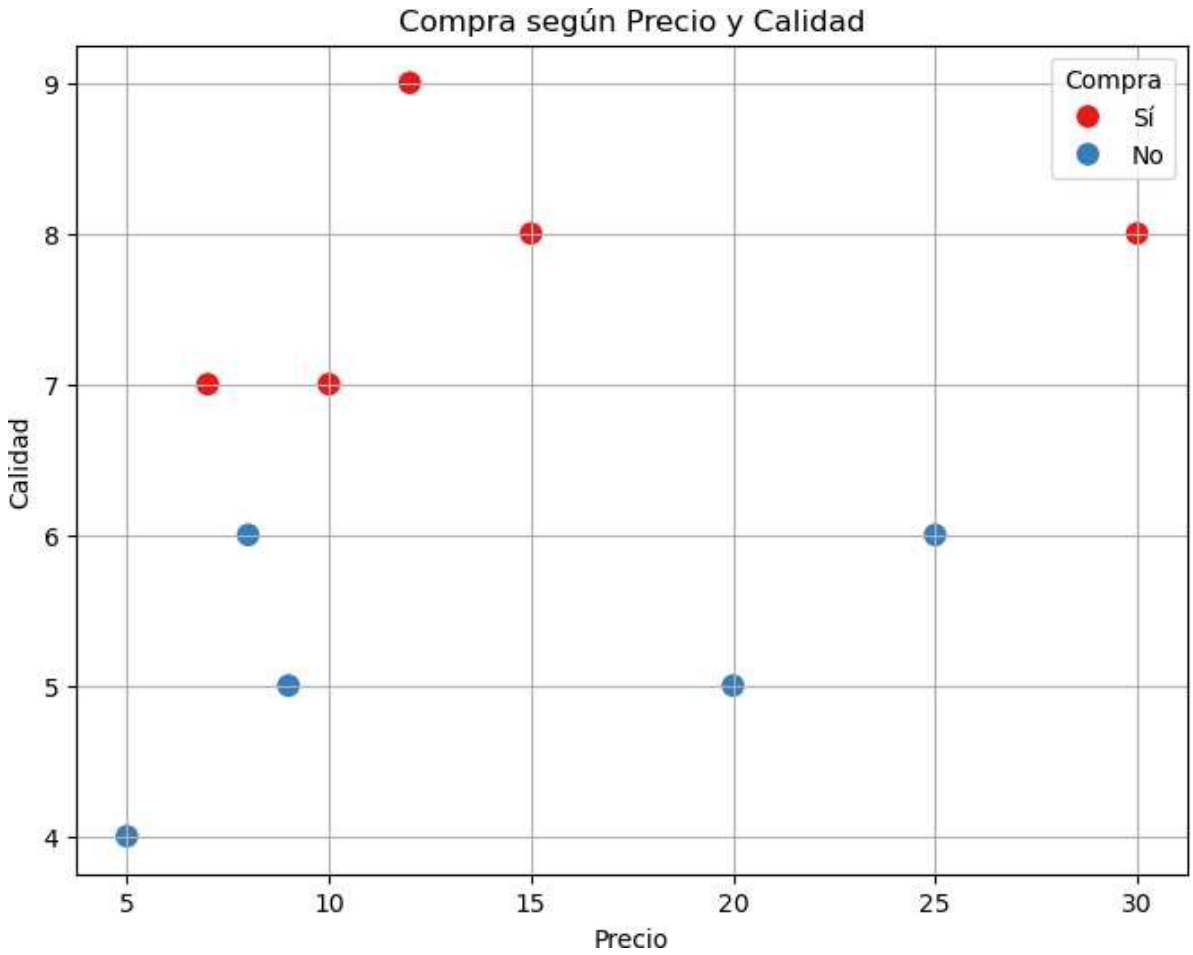
0	1.00	1.00	1.00	1
1	1.00	1.00	1.00	2
accuracy			1.00	3
macro avg	1.00	1.00	1.00	3
weighted avg	1.00	1.00	1.00	3

```
Score: 1.00
```

1.5 Visualización

```
import seaborn as sns

# Para visualizar la separación

plt.figure(figsize=(8,6))
sns.scatterplot(x='Precio', y='Calidad', hue='Compra', data=df,
```

```
palette='Set1', s=100)
plt.title('Compra según Precio y Calidad')
plt.xlabel('Precio')  plt.ylabel('Calidad')
plt.grid(True)
plt.show()
```



1.6 Interpretación

```
# Interpretación de coeficientes

coef_df = pd.DataFrame(lda.coef_, columns=X.columns) print("\nCoeficientes del modelo (mayor valor = más influencia en la decisión):")

print(coef_df.T)
```

```
Coeficientes del modelo (mayor valor = más influencia en la decisión):
Precio
```

```
Descuento  1.058287
Calidad    4.374470
```

1.7 Nuevos productos

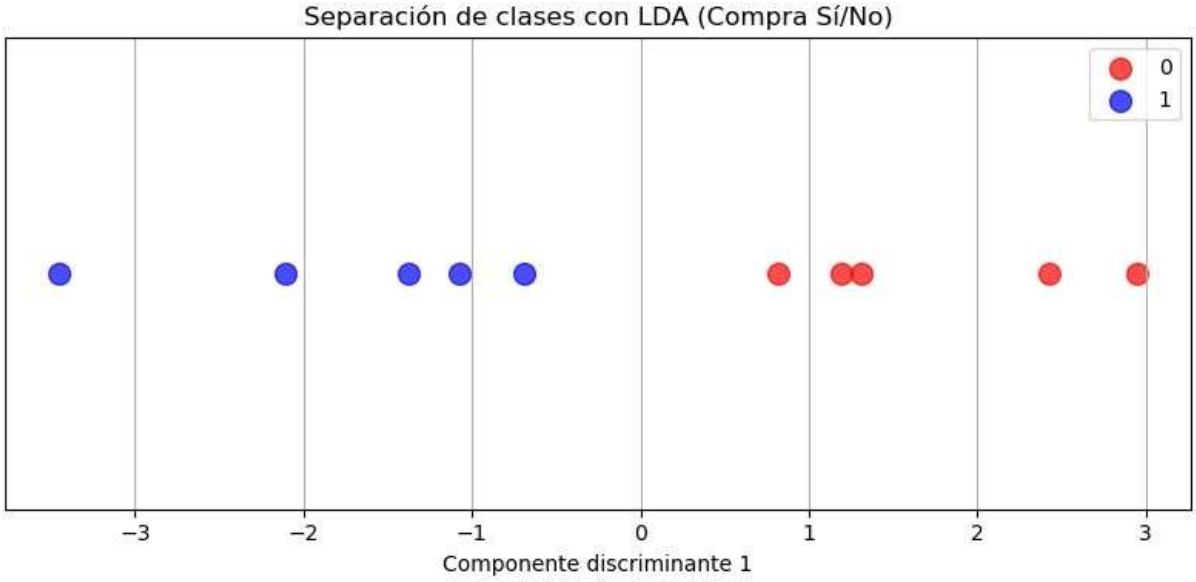
```
# Probabilidades a posteriori para X_test
probas = lda.predict_proba(X_test)

# Primera fila de probabilidades print("Probabilidades a posteriori: ") print(pro-
bas[0])

# Clasificación de nuevos productos
nuevos_productos = pd.DataFrame({ 'Precio': [5, 26, 11],
'Descuento': [2, 0, 1],
'Calidad': [8, 7, 3]
})

predicciones_nuevas = lda.predict(nuevos_productos) print("\nPredicción para nuevos productos:")
for i, pred in enumerate(predicciones_nuevas): clase = "Sí" if pred == 1 else "No"
print(f"Producto {i+1}: ¿Compra? → {clase}")
```

```
Probabilidades a posteriori: [0.0084219 0.9915781]

Predicción para nuevos productos: Producto 1: ¿Compra?
→ Sí Producto 2: ¿Compra? → No Producto 3: ¿Compra?
→ No
```

1.8 Gráfico

```
# Codificamos la variable dependiente (Sí = 1, No = 0)

le = LabelEncoder()
y_encoded = le.fit_transform(y)

# Aplicamos LDA

X_lda = lda.fit_transform(X, y_encoded)

# Graficamos los datos transformados

plt.figure(figsize=(8, 4))

for clase, color, label in zip(np.unique(y_encoded), ['red', 'blue'],
le.classes_):
plt.scatter(
```

```
label=label, c=color,
alpha=0.7, s=100
)

plt.title('Separación de clases con LDA (Compra Sí/No)') plt.xla-
bel('Componente discriminante 1')
plt.yticks([]) # Ocultamos eje Y

plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```



a) ¿Qué variables contribuyen más a la probabilidad de compra y cómo?

La calidad es la varialbe que más contribuye, cuatro veces más que el descuento. El precio, por tener coeficiente negativo, contrubuye a la probabilidad de "no compra".

b) Haz un gráfico para visualizar la segmentación. 1.8
c) ¿El modelo categoriza bien los nuevos productos? Sí, La matriz de confusión es una matriz diagonal.
d) ¿Existe un umbral de precio a partir del cual la probabilidad de compra disminuye drásticamente?

No, vemos en el gráfico Calidad vs Precio que el producto con precio 30 sí fue comprado.

Ejercicio Resuelto

Una empresa de retail desea predecir si un consumidor comprará o no un producto en función de tres características clave del producto: el precio, el porcentaje de descuento ofrecido y la percepción de calidad. Para ello, se ha recopilado un conjunto de datos históricos que registra, para cada producto ofrecido, el precio final, el tipo de descuento aplicado(0 = sin descuento, 1 = de 5 a 10%, 2 = de 15 a 25%), la calidad percibida (en una escala de 1 a 10) y si finalmente fue comprado (Sí) o no (No). Se desea aplicar Análisis Discriminante para construir un modelo de clasificación que permita predecir la decisión de compra de futuros productos con base en estas variables.

1.0 Datos

```
import pandas as pd im-
port numpy as np
from sklearn.discriminant_analysis import
QuadraticDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score, clas-
sification_report
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder

# Crear un dataset

data = {
'Precio': [10, 15, 8, 12, 20, 5, 25, 7, 30, 9],
'Descuento': [1, 2, 0, 2, 0, 1, 0, 2, 1, 3],
'Calidad': [7, 8, 6, 9, 5, 4, 6, 7, 8, 5],
'Compra': ['Sí', 'Sí', 'No', 'Sí', 'No', 'No', 'No', 'Sí', 'Sí', 'No']
}

df = pd.DataFrame(data)
```

1.1 Preparar datos

```
# Codificar variable dependiente

le = LabelEncoder()
df['Compra_encoded'] = le.fit_transform(df['Compra']) # 'Sí' → 1, 'No' → 0

X = df[['Precio', 'Descuento', 'Calidad']]
y = df['Compra_encoded']

# Dividir en entrenamiento y prueba
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)
```

1.2 Entrenar modelo

```
# Entrenar modelo QDA

qda = QuadraticDiscriminantAnalysis()
qda.fit(X_train, y_train)
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\ discriminant_analy-
sis.py:1024: LinAlgWarning: The covariance matrix of class 1 is not full
rank. Increasing the value of parameter
`reg_param` might help reducing the collinearity.
  warnings.warn(
```

```
QuadraticDiscriminantAnalysis()
```

1.3 Predicciones

```
# Predicciones

y_pred = qda.predict(X_test)
```

1.4 Evaluación

```
# Evaluación
print("Matriz de Confusión:") print(confusion_matrix(y_test, y_pred))

print("\nReporte de Clasificación:") print(classification_report(y_test, y_pred))

print(f"Score: {accuracy_score(y_test, y_pred):.2f}")
```

```
Matriz de Confusión:
[[1 0]
 [2 0]]

Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.33      1.00      0.50         1
           1       0.00      0.00      0.00         2

    accuracy                           0.33         3
   macro avg       0.17      0.50      0.25         3
weighted avg       0.11      0.33      0.17         3

Score: 0.33

C:\Users\San\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
```

```
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
C:\Users\San\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1565: UndefinedMetricWarning: Precision is ill- de-
fined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
C:\Users\San\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1565: UndefinedMetricWarning: Precision is ill- de-
fined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
```

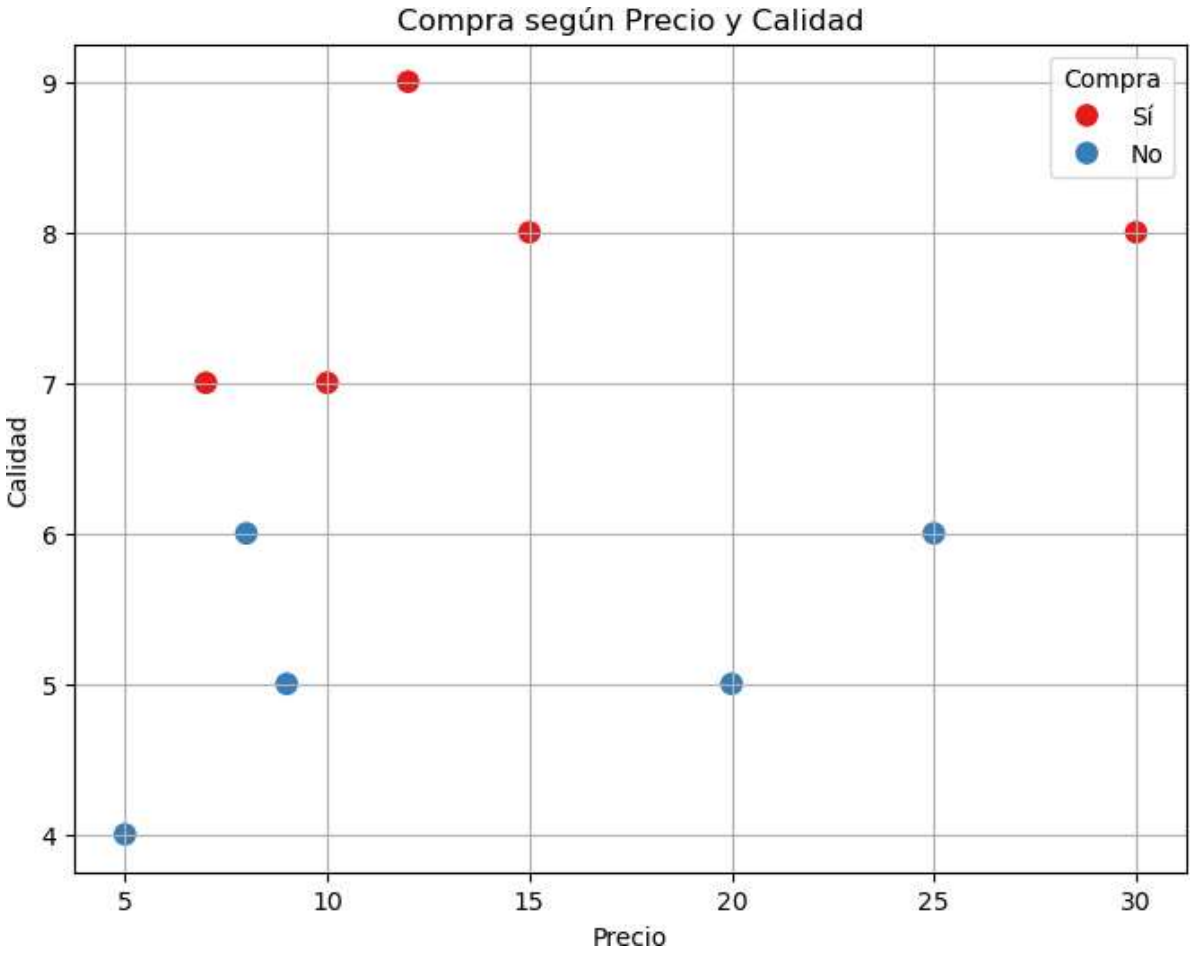
1.5 Interpretación

```
# Interpretación de coeficientes print("Clases:", qda.classes_)
print("Medias por clase:\n", qda.means_)
```

```
Clases: [0 1] Medias por
clase:
 [[15.5          0.75          5.5          ]
  7.66666667]]
```

1.6 Visualización

```
import seaborn as sns

# Para visualizar la separación

plt.figure(figsize=(8,6))
sns.scatterplot(x='Precio', y='Calidad', hue='Compra', data=df,
palette='Set1', s=100)
plt.title('Compra según Precio y Calidad')
plt.xlabel('Precio')  plt.ylabel('Calidad')
plt.grid(True)
plt.show()
```



1.7 Nuevos productos

```
# Probabilidades a posteriori para X_test

probas = qda.predict_proba(X_test)

# Primera fila de probabilidades print("Probabilidades a posteriori: ")
print(probas[0])

# Clasificación de nuevos productos

nuevos_productos = pd.DataFrame({
'Precio': [5, 26, 11],
'Descuento': [2, 0, 1],
'Calidad': [8, 7, 3]
})

predicciones_nuevas = qda.predict(nuevos_productos) print("\nPre-
dicción para nuevos productos:")
for i, pred in enumerate(predicciones_nuevas):
```

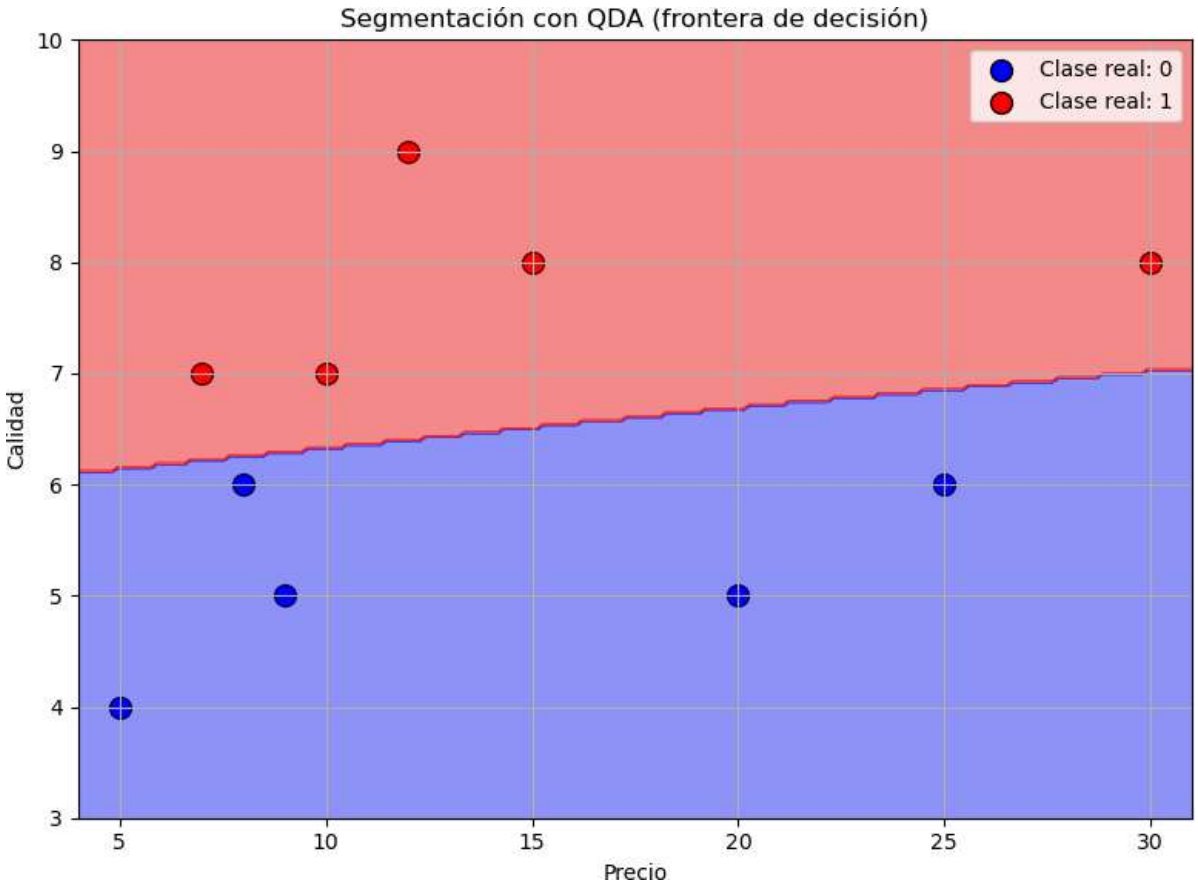
```
clase = "Sí" if pred == 1 else "No" print(f"Pro-
ducto {i+1}: ¿Compra? → {clase}")

Probabilidades a posteriori:
[1. 0.]

Predicción para nuevos productos:
Producto 1: ¿Compra? → No Pro-
ducto 2: ¿Compra? → No Producto
3: ¿Compra? → No
```

1.8 Gráfico

```
# Codificamos la variable dependiente (Sí = 1, No = 0)
le = LabelEncoder()
y_encoded = le.fit_transform(y)

# Variables a graficar
X = df[['Precio', 'Calidad']]

# Aplicamos QDA
X_qda = qda.fit(X, y_encoded)

# Crear una malla para graficar la frontera
x_min, x_max = X['Precio'].min() - 1, X['Precio'].max() + 1
y_min, y_max = X['Calidad'].min() - 1, X['Calidad'].max() + 1 xx, yy = np.meshgrid(np.linspace(x_min,
x_max, 200),
np.linspace(y_min, y_max, 200))

# Predecimos sobre la malla
Z = qda.predict(np.c_[xx.ravel(), yy.ravel()]) Z = Z.reshape(xx.shape)

# Gráfico de fondo: regiones
plt.figure(figsize=(8,6))
plt.contourf(xx, yy, Z, alpha=0.3, cmap='coolwarm')

# Obtenemos clases codificadas
clases = np.unique(y_encoded)

# Mapear colores a clases
colormap = {clase: color for clase, color in zip(clases, ['blue', 'red'])}

# Puntos reales
for clase in clases: plt.scatter(
df[y_encoded == clase]['Precio'], df[y_encoded == clase]['Calidad'], c=color-
map[clase],
```

```
edgecolor='k', s=100,
label=f'Clase real: {le.inverse_transform([clase])[0]}'
)

# Fondo (regiones de decisión)
from matplotlib.colors import ListedColormap

# Mismo orden de colores usados en puntos
cmap_background = ListedColormap([colormap[clase] for clase in clases])
plt.contourf(xx, yy, Z, alpha=0.3, cmap=cmap_background)

plt.title('Segmentación con QDA (frontera de decisión)') plt.xlabel('Precio')
plt.ylabel('Calidad')                    plt.legend()
plt.grid(True) plt.tight_layout() plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\utils\ validation.py:2739: UserWarning: X does not have valid feature names, but QuadraticDiscriminantAnalysis was fitted with feature names
  warnings.warn(
```



a) ¿Qué variables contribuyen más a la probabilidad de compra y cómo? Calidad, como se ve en los gráficos.
b) Haz un gráfico para visualizar la segmentación. 1.8
c) ¿El modelo categoriza bien los nuevos productos?

No, los categoriza todos como "No compra" También podemos verlo en la matriz de confusión, que tiene una columna de 0. Lo vemos además en el Score: solo acierta el 33% de las veces.

d) ¿Existe un umbral de precio a partir del cual la probabilidad de compra disminuye drásticamente?

No, el producto más caro si fue comprado.

Ejercicio de Análisis Factorial

Una empresa de telecomunicaciones busca segmentar a sus usuarios en base a **comportamientos observados** y no sus preferencias personales.

Variables medidas en 5 usuarios:

Usuario	**Minutos al mes**	**Número de llamadas**	**Datos móviles (GB)**
U1	800	200	3.5
U2	750	180	4.0
U3	300	50	12.0
U4	280	45	11.0
U5	700	190	2.0

Preguntas

a) Identifica los factores que extraídos.
b) ¿Cuál es la variable más representativa? ¿Qué porcentaje de la varianza total explica?
c) ¿Qué segmentos de usuarios se pueden definir a partir de los factores?
d) ¿Qué tipo de oferta o plan conviene diseñar para cada segmento identificado?

Ejercicio Resuelto

Una empresa de telecomunicaciones busca segmentar a sus usuarios en base a comportamientos observados y no sus preferencias personales.

1.0 Datos

```
import pandas as pd
from factor_analyzer import FactorAnalyzer
from factor_analyzer.factor_analyzer import calculate_kmo, calcu-
late_bartlett_sphericity
import matplotlib.pyplot as plt

# Datos

data = {
'Minutos': [800, 750, 300, 280, 700],
'Llamadas': [200, 180, 50, 45, 190],
'Datos_móviles': [3.5, 4.0, 12.0, 11.0, 2.0]
}
```

1.1 Verificación de adecuación

```
# KMO

kmo_all, kmo_model = calculate_kmo(df)
print(f"KMO general: {kmo_model:.2f}")

# Bartlett

chi_square_value, p_value = calculate_bartlett_sphericity(df)
print(f"Bartlett Test: chi2 = {chi_square_value:.2f}, p =
{p_value:.4f}")

KMO general: 0.55
```

KMO > 0.5, pero bajo, lo damos por válido. Esto nos informa de que el análisis factorial puede no ser la mejor opción.

Barlett $p < 0.05$, datos adecuados para este análisis.

1.2 Análisis factorial

```
fa = FactorAnalyzer(n_factors=1, rotation=None)
fa.fit(df)
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\utils\ depreca-
tion.py:151: FutureWarning: 'force_all_finite' was renamed to 'en-
sure_all_finite' in 1.6 and will be removed in 1.8.
   warnings.warn(

FactorAnalyzer(n_factors=1, rotation=None, rotation_kwargs={})
```

1.3 Matriz de correlación

```
correlation_matrix = df.corr() print(correlation_matrix)
```

	Minutos	Llamadas	Datos_móviles
Minutos	1.000000	0.992518	-0.956441
Llamadas			
Datos_móviles	-0.956441	-0.981829	

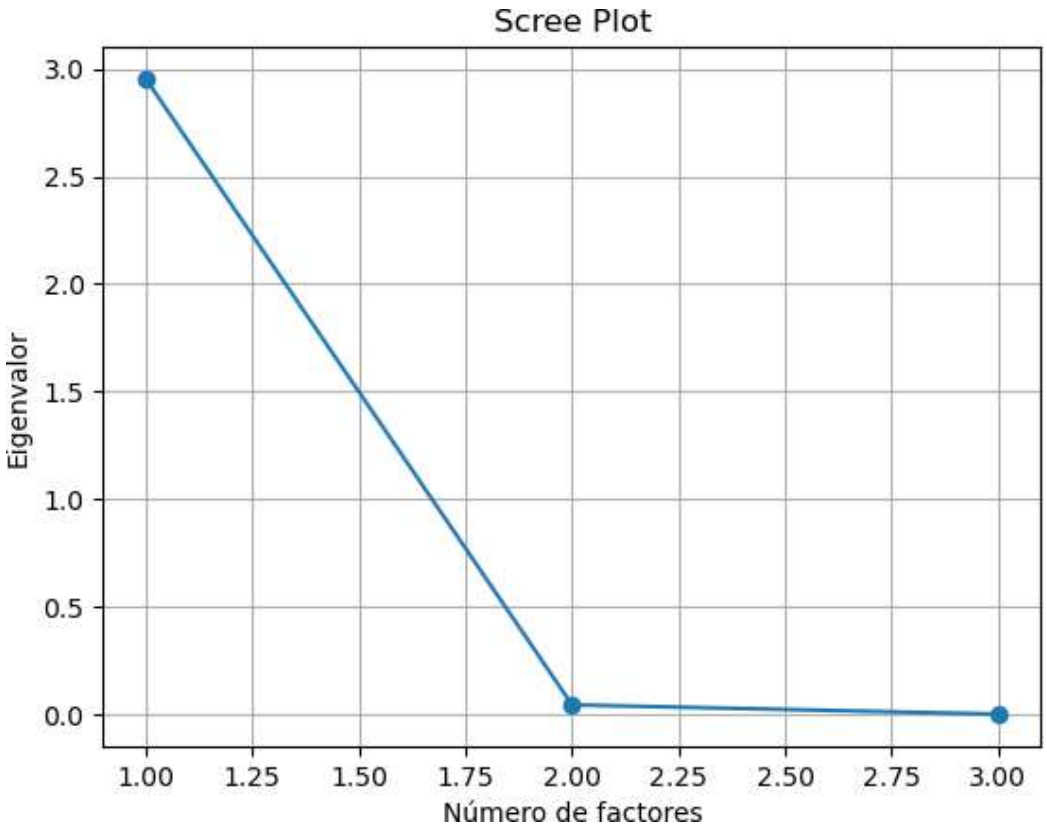
1.4 Eigenvalues

```
ev, _ = fa.get_eigenvalues() print("\nEigen-
valores:", ev)


Eigenvalores: [2.95393671e+00 4.44543673e-02 1.60892202e-03]
```

Extraemos 1 factor.

1.5 Scree plot

```
plt.plot(range(1, len(ev)+1), ev, marker='o')
plt.title('Scree Plot')
plt.xlabel('Número de factores')
plt.ylabel('Eigenvalor')
plt.grid(True)
plt.show()
```



Extraemos 1 factor.

1.6 Cargas factoriales y Comunalidades

```
#Carga factorial

loadings = fa.loadings_

print("\nCargas factoriales:\n", pd.DataFrame(loadings, index=df.columns))

# Comunalidades

comunalidades = pd.DataFrame(fa.get_communalities(), index=df.columns, columns=["Comunalidad"])
```

```
Cargas factoriales:

Minutos          -0.985799
Llamadas
Datos_móviles     0.975081

                 Comunalidad
Minutos              0.971800
```

```
Llamadas Da-        1.004175
tos_móviles
```

1.7 Scores

```
scores = fa.transform(df)
print("\nPuntuaciones del factor por usuario:\n", pd.DataFrame(scores,
index=df.index))


Puntuaciones del factor por usuario:
             0
0 -1.073455
1 -0.426111
2  0.976046
3  1.435740
4 -0.912219

C:\Users\San\anaconda3\Lib\site-packages\sklearn\utils\ depreca-
tion.py:151: FutureWarning: 'force_all_finite' was renamed to 'en-
sure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
```

a) Identifica los factores que extraídos.

Extraemos un único factor. Analizando las cargas factoriales vemos que:

- cuando los scores son bajos, el usuario tiende a usar más minutos y llamadas
- cuando los scores son altos, el usuario tiende a usar más datos móviles
- podemos decir también que un score en torno a 0 es de un usuario equilibrado De esta forma, podemos segmentar el mercado en tres tipos de usuarios distintos.

En base a esto, podemos denominar a este factor: "Preferencia de mensajes escritos vs Preferencia de llamadas"

b) ¿Cuál es la variable más representativa? ¿Qué porcentaje de la varianza total explica?

No hay una variable más representativa que otras. Las tres tienen una conmunalidad similar.

c) ¿Qué segmentos de usuarios se pueden definir a partir de los factores? Tres tipos de usuario:

- Usuarios que prefieren usar datos móviles
- Usuarios que prefieren llamar
- Usuarios que equilibran ambos medios de comunicación

d) ¿Qué tipo de oferta o plan conviene diseñar para cada segmento identificado? Siguiendo el orden de la anterior respuesta:

- Tarifa de datos ilimitados
- Tarifa de minutos gratis
- Tarifa plana

Ejercicio de Análisis de Correspondencia

Una empresa quiere analizar las **preferencias de distintos grupos de clientes** (segmentados por edad o tipo) respecto a **cuatro productos distintos**. Para ello, ha recopilado una tabla con las **frecuencias de compra** de cada producto en función del segmento:

- Jóvenes
- Adultos
- Mayores
- Empresas

El objetivo del análisis es **identificar patrones de asociación entre productos y segmentos**, descubrir **grupos de productos similares** y **segmentos que comparten comportamientos**, con el fin de **orientar futuras estrategias de marketing, distribución y comunicación**.

	Jóvenes	Adultos	Mayores	Empresas
Producto A	40	5	3	2
Producto B	10	20	5	3
Producto C	5	10	40	5
Producto D	2	5	5	30

Preguntas

a) ¿Hay productos que los jóvenes prefieren más que otros?

b) ¿Qué productos podrían considerarse de nicho por su uso concentrado en un solo grupo?

c) ¿Existen productos con un perfil más generalista o equilibrado?

d) ¿Qué segmentos podrían representar oportunidades de marketing cruzado?

e) ¿Qué representa cada componente?

Ejercicio Resuelto

Una empresa quiere analizar las preferencias de distintos grupos de clientes (segmentados por edad o tipo) respecto a cuatro productos distintos. El objetivo del análisis es identificar patrones de asociación entre productos y segmentos, descubrir grupos de productos similares y segmentos que comparten comportamientos, con el fin de orientar futuras estrategias de marketing, distribución y comunicación.

1.0 Datos

```
import pandas as pd import numpy as np
import matplotlib.pyplot as plt import seaborn as sns
from sklearn.preprocessing import StandardScaler
import prince   # Librería para Análisis de Correspondencias

# Tabla de contingencia
data = {
'Producto A': [40, 5, 3, 2],
'Producto B': [10, 20, 5, 3],
'Producto C': [5, 10, 40, 5],
'Producto D': [2, 5, 5, 30]
}
clientes = ['Jóvenes', 'Adultos', 'Mayores', 'Empresas']

df = pd.DataFrame(data, index=clientes) print(df)
```

	Producto A	Producto B	Producto C	Producto D
Jóvenes	40	10	5	2
Adultos	5	20	10	5
Mayores	3	5	40	5
Empresas	2	3	5	30

1.1 Análisis de Correspondencias

```
# Crear el modelo MCA con sklearn como motor

mca = prince.MCA(
n_components=2, en-
gine='sklearn', ran-
dom_state=42
)

# Ajustar el modelo
```

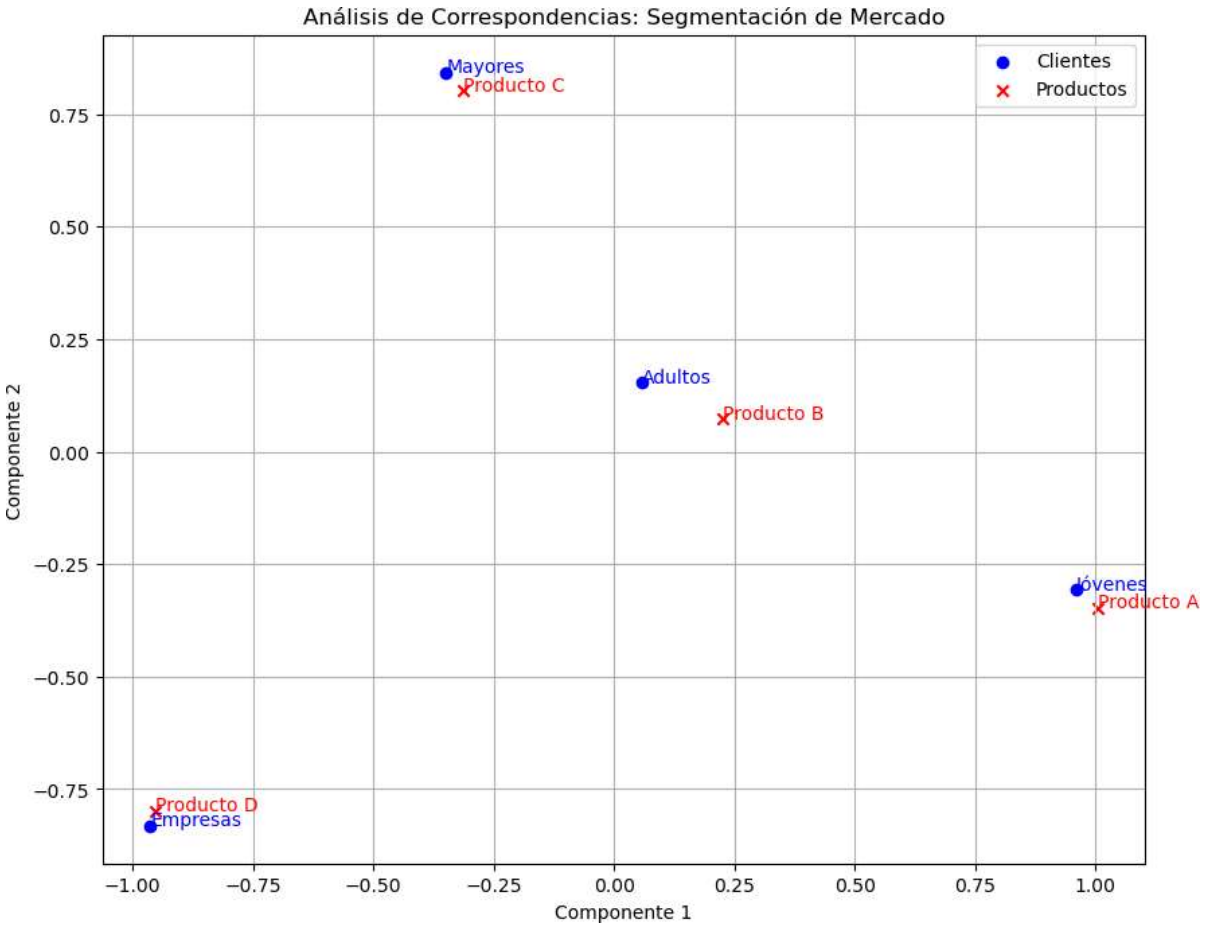
```
mca = mca.fit(df)

# Obtener coordenadas de individuos (filas)

row_coords = mca.row_coordinates(df)

# Obtener coordenadas de categorías (columnas)

col_coords = mca.column_coordinates(df)

print("Coordenadas de los clientes:\n", row_coords) print("\nCoordenadas de los productos:\n", col_coords)
```

```
Coordenadas de los clientes:

0                                1
Jóvenes
Adultos
                  0.959852 -0.307181
Mayores
Empresas -0.963334 -0.831458
```

```
Coordenadas de los productos:

0                                   1

Producto A   1.003580 -0.347276

Producto B   0.226458   0.072995
```

1.2 Gráfica

```
# Graficar resultados

plt.figure(figsize=(10, 8))

# Filas

plt.scatter(row_coords[0], row_coords[1], c='blue', label='Clientes')
for i, txt in enumerate(df.index):
plt.annotate(txt, (row_coords.iloc[i, 0], row_coords.iloc[i, 1]),
color='blue')

# Columnas

plt.scatter(col_coords[0], col_coords[1], c='red', label='Productos',
marker='x')
for i, txt in enumerate(df.columns):
plt.annotate(txt, (col_coords.iloc[i, 0], col_coords.iloc[i, 1]),
color='red')

plt.title('Análisis de Correspondencias: Segmentación de Mercado')
plt.xlabel('Componente 1')
plt.ylabel('Componente 2') plt.leg-
end()
```



1.3 Contribuciones

```
col_coords = mca.column_coordinates(df)

# Cuadrados de las coordenadas

col_coords_sq = col_coords**2

# Contribución de cada categoría a cada dimensión (en %) dim1_contrib = 100 * col_coords_sq[0] /
col_coords_sq[0].sum() dim2_contrib = 100 * col_coords_sq[1] / col_coords_sq[1].sum()

# Mostrar como tabla en vez de usar print con formato de número

contrib_df = pd.DataFrame({
'Contribución Dim 1 (%)': dim1_contrib, 'Contribu-
ción Dim 2 (%)': dim2_contrib
})

print(contrib_df.round(2))

# Varianza de cada dimension total_var_dim1 =
sum(col_coords[0]**2) total_var_dim2 =
```

```
print(f"\nVarianza aproximada Dim 1: {total_var_dim1:.4f}") print(f"Varianza aproxi-
mada Dim 2: {total_var_dim2:.4f}")
```

	Contribución Dim 1 (%)	Contribución Dim 2 (%)
Producto A	48.84	8.58
Producto B	2.49	0.38
Producto C	4.78	45.73
Producto D	43.89	45.31

```
Varianza aproximada Dim 1: 2.0620
```

a) ¿Hay productos que los jóvenes prefieren más que otros? Sí, el producto A.

b) ¿Qué productos podrían considerarse de nicho por su uso concentrado en un solo grupo? Podríamos considerar de nicho:

- Jóvenes y el producto A
- Mayores y el producto C
- Empresas y el producto D

c) ¿Existen productos con un perfil más generalista o equilibrado? Adultos y el producto B están en el centro del gráfico. Puede entenderse:

- que los Adultos compras más el producto B
- que los Adultos compran todos los productos por igual
- y que el producto B es comprado por igual por todos los segmentos de clientes

Ejercicio de Análisis Clúster – K-means

Un centro comercial quiere segmentar a sus clientes para ofrecer promociones más personalizadas. Para ello, ha recolectado datos de 200 clientes con las siguientes variables:

- **Edad** (años)
- **Ingreso mensual** (en miles de euros)
- **Puntaje de gasto** (del 1 al 100, donde 100 indica que el cliente gasta mucho en el centro comercial)

Aplica **K-Means** para identificar patrones de comportamiento y dividir a los clientes en grupos significativos.

Ingreso	PuntajeGasto	Edad
3172.51	2	41
4635.48	58	59
3037.33	55	36
2115.85	81	31

Preguntas

a) ¿Qué valor de **k** sugiere el **método del codo**?

b) ¿Cuál es el valor del **coeficiente de silueta** para k = 2, 3, 4?

c) ¿Qué valor de **k** parece generar los grupos más separados y compactos?

d) ¿Puedes describir brevemente el perfil de cada grupo?

e) ¿Qué tipo de estrategia de marketing aplicarías para cada segmento?

f) ¿Qué clúster representa a los clientes más valiosos para la empresa?

g) ¿Cómo podrías usar esta segmentación para personalizar promociones?

h) ¿Los clústeres se visualizan claramente en un gráfico de dispersión?

i) ¿Se superponen los clústeres en algunas combinaciones de variables?

Ejercicio Resuelto

Un centro comercial quiere segmentar a sus clientes para ofrecer promociones más personalizadas. Para ello, ha recolectado datos de 200 clientes. Tu tarea es aplicar K-Means para identificar patrones de comportamiento y dividir a los clientes en grupos significativos.

1.0 Datos

```
import numpy as np import pan-
das as pd

import matplotlib.pyplot as plt import seaborn as
sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler from sklearn.metrics
import silhouette_score

np.random.seed(321)

# Datos

n = 200

ingreso = np.random.normal(3000, 1000, n)                    # en euros pun-
taje_gasto = np.random.randint(1, 101, n)                    # entre 1 y 100 edad =
np.random.randint(18, 65, n)

df = pd.DataFrame({ 'Ingreso': ingreso,

'PuntajeGasto': puntaje_gasto, 'Edad': edad

})
d
```

	Ingreso	PuntajeGasto	Edad
0	3172.519469	2	41
1	4635.482532	58	59
2	3037.336403	55	36

1.1 Estandarización

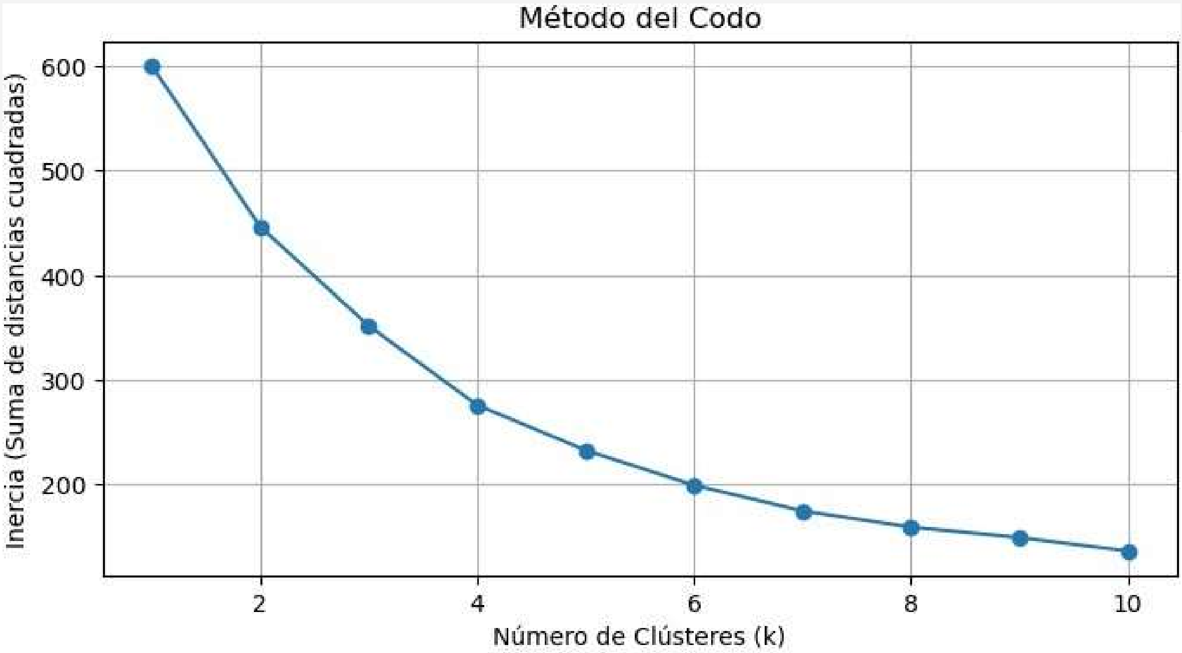
```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df)
```

1.2 Gráfico de codo

```
inertias = []
K_range = range(1, 11)

for k in K_range:
kmeans = KMeans(n_clusters=k, random_state=42, n_init=10) kmeans.fit(X_scaled)
inertias.append(kmeans.inertia_)

plt.figure(figsize=(8, 4)) plt.plot(K_range, inertias, marker='o') plt.title("Método del Codo") plt.xlabel("Número de Clústeres (k)")
plt.ylabel("Inercia (Suma de distancias cuadradas)") plt.grid(True)
plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
```

Windows with MKL, when there are less chunks than available threads. You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(

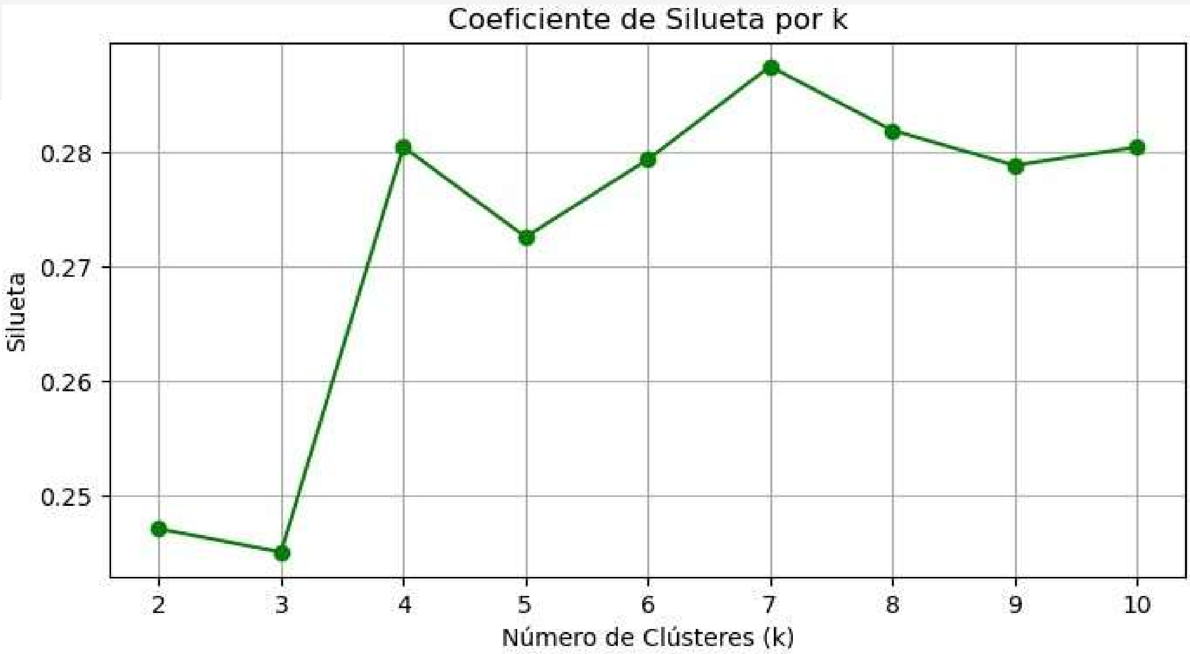


1.3 Coeficiente silueta

```
# 🧠 Coeficiente de silueta para distintos k
silhouette_scores = [] K_range_eval = range(2, 11)

for k in K_range_eval:
kmeans = KMeans(n_clusters=k, random_state=42, n_init=10) labels = kmeans.fit_predict(X_scaled)
score = silhouette_score(X_scaled, labels) silhouette_scores.append(score)

plt.figure(figsize=(8, 4))
plt.plot(K_range_eval, silhouette_scores, marker='o', color='green') plt.title("Coeficiente de Silueta por k")
plt.xlabel("Número de Clústeres (k)") plt.ylabel("Silueta")
plt.grid(True) plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there
are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there
are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there
are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there
are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there
are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```



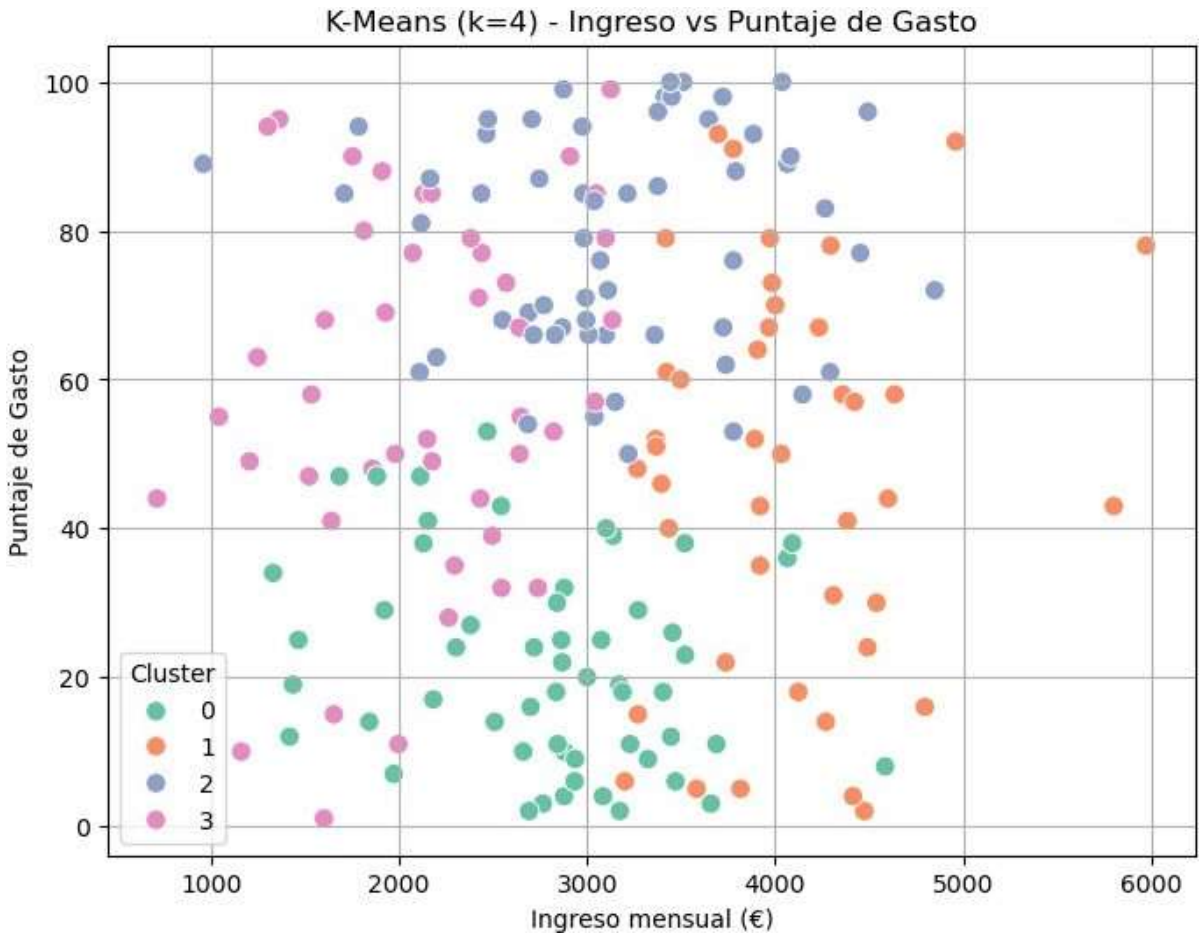
1.4 K-means final

```
# Elegimos k=4

k_final = 4
kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=10)
df['Cluster'] = kmeans_final.fit_predict(X_scaled)
```

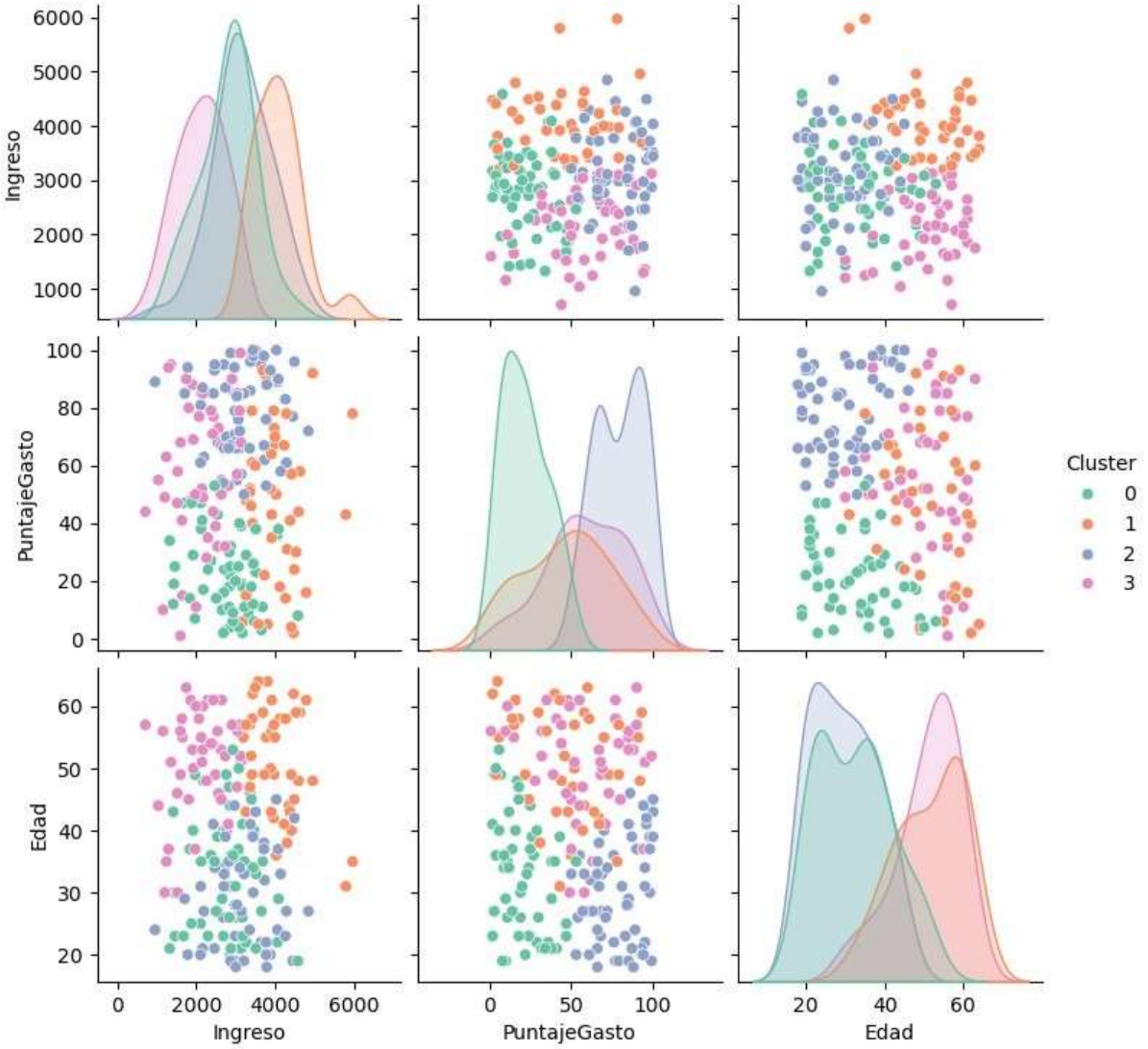
```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```

1.5 Gráfico de dispersión

```
# Ingreso vs. Puntaje de Gasto

plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='Ingreso', y='PuntajeGasto', hue='Cluster',
palette='Set2', s=70)
plt.title(f"K-Means (k={k_final}) - Ingreso vs Puntaje de Gasto")
plt.xlabel("Ingreso mensual (€)")
plt.ylabel("Puntaje de Gasto")
plt.legend(title="Cluster")
plt.grid(True)
plt.show()
```




```
# Ver agrupamiento

sns.pairplot(df, hue='Cluster', palette='Set2')
plt.show()
```

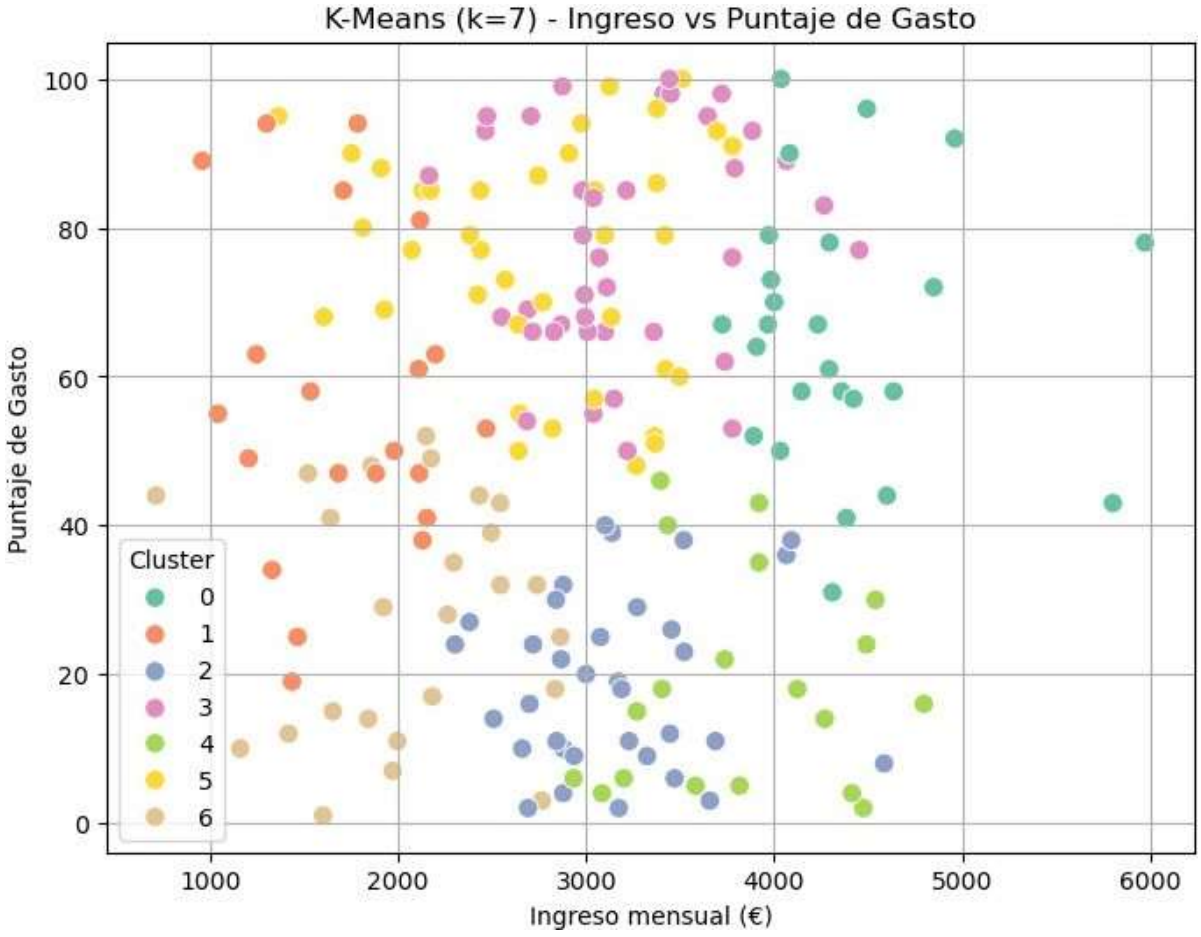


1.6 Repetimos con k = 7

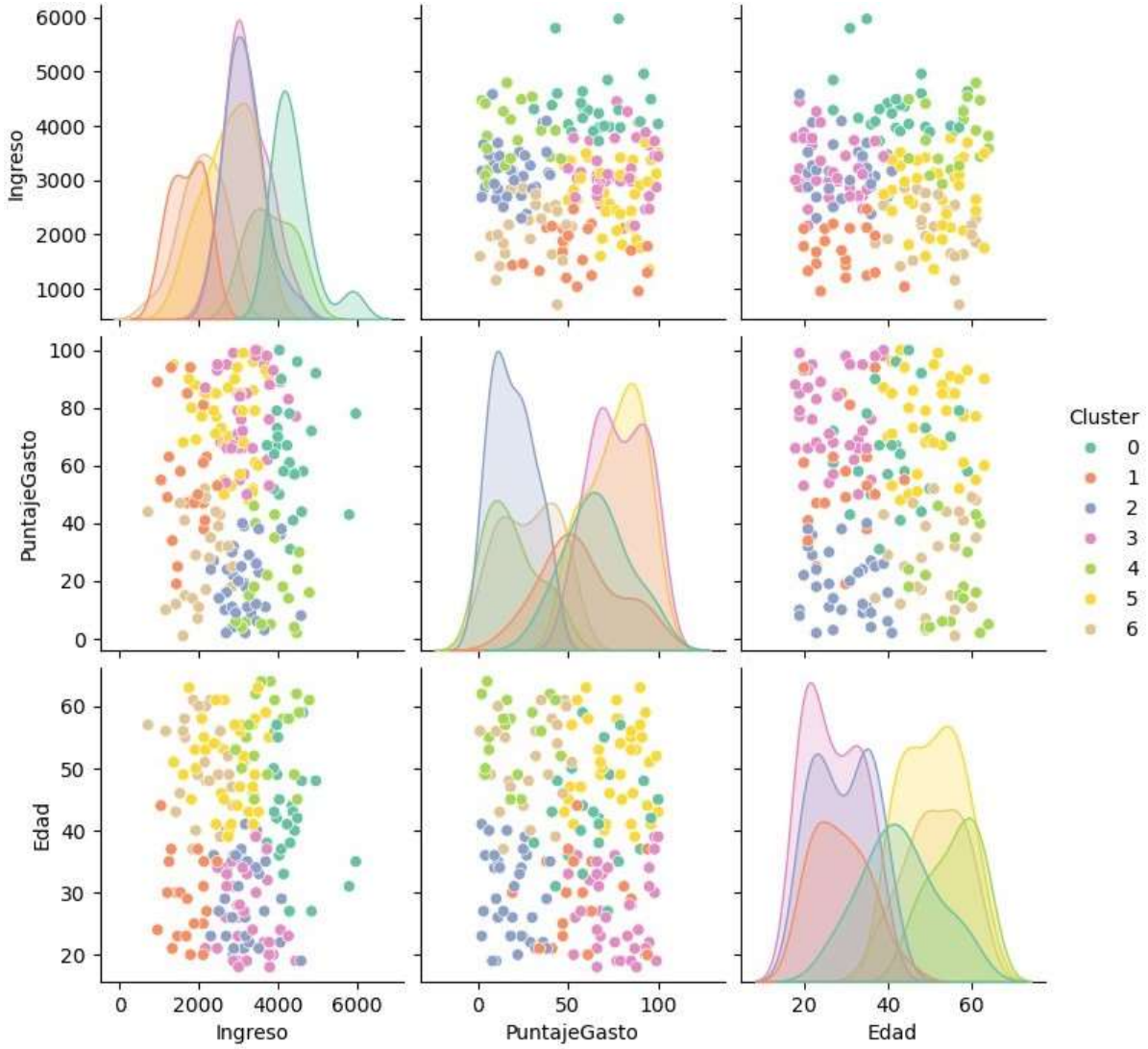
```
# Elegimos k=7

k_final = 7
kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=10)
df['Cluster'] = kmeans_final.fit_predict(X_scaled)

# Ingreso vs. Puntaje de Gasto

plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='Ingreso', y='PuntajeGasto', hue='Cluster',
palette='Set2', s=70)
plt.title(f"K-Means (k={k_final}) - Ingreso vs Puntaje de Gasto")
plt.xlabel("Ingreso mensual (€)")
plt.ylabel("Puntaje de Gasto") plt.legend(ti-
```

```
plt.grid(True)
plt.show()
```

```
C:\Users\San\anaconda3\Lib\site-packages\sklearn\cluster\
_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on
Windows with MKL, when there are less chunks than available threads.
You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
   warnings.warn(
```




```
# Ver agrupamiento

sns.pairplot(df, hue='Cluster', palette='Set2')
plt.show()
```



a) ¿Qué valor de k sugiere el método del codo?

No se aprecia claramente un codo, pero podemos aventurarnos con k=4, ya que la pendiente disminuye.

b) ¿Cuál es el valor del coeficiente de silueta para k = 2, 3, 4?

En k = 2, 3 es menor que 0.25 En k = 4 es 0.28

c) ¿Qué valor de k parece generar los grupos más separados y compactos? k = 7

d) ¿Puedes describir brevemente el perfil de cada grupo? Siendo k = 4

- Jóvenes de ingreso medio y gasto bajo [verde]
- Jóvenes de ingreso medio y gasto alto [azul]
- Mayores de ingreso alto y gasto medio [rojo]
- Mayores de ingreso bajo y gasto medio [rosa]

e) ¿Qué tipo de estrategia de marketing aplicarías para cada segmento? Siguiendo el orden de la pregunta anterior:

- Ofertas de entrada, campañas de fidelización y muostrar el valor y beneficios concretos del producto.
- Campañas experienciales, exclusividad asequible, personalización del producto o servicio, up-selling y cross-selling.

- Servicio premium, email marketing segmentado y contenido informativo, productos complementarios de valor añadido y destacar beneficios a largo plazo, garantías y prestigio de marca.
- Promociones por ahorro o conveniencia, facilidades de pago y programas de fidelidad.

f) ¿Qué clúster representa a los clientes más valiosos para la empresa?

Los jóvenes de ingreso medio y gasto alto, pues tienen una mayor proyección y se podría esperar que con el paso del tiempo sus ingresos aumentaran como lo haría también su gasto.

g) ¿Cómo podrías usar esta segmentación para personalizar promociones?

Se podrían usar las estrategias de marketing que ya se explicaron anteriormente y llevarlas a cabo coordinadamente con los eventos anuales como pueden ser: el inicio de las vacaciones, la navidad y la temporada de rebajas.

h) ¿Los clústeres se visualizan claramente en un gráfico de dispersión?

No

i) ¿Se superponen los clústeres en algunas combinaciones de variables?

Sí, en el punto medio de las tres variables hay clientes que podrían pertenecer a cualquiera de las clases.

Ejercicio de Análisis Clúster – Ward

Una agencia de viajes online quiere segmentar a sus clientes para enviarles ofertas personalizadas según su perfil de viajero. Cuenta con información de 100 clientes, con las siguientes variables:

- Número de viajes internacionales por año
- Gasto promedio por viaje (en euros)
- Días promedio por viaje
- Tipo de viaje preferido (codificado como: 0 = Negocios, 1 = Placer)

Utilizar el método jerárquico de Ward para agrupar a los clientes en segmentos, visualizar la estructura de agrupación con un dendrograma, y definir el número óptimo de clústeres.

ViajesAnuales	GastoPromedio	DiasPromedio	TipoViaje
4	1023.190573	4.359534	0
1	1454.880629	10.662918	1
3	1307.104646	9.358880	1
3	992.127121	6.061649	1

Ejercicio Resuelto

Una agencia de viajes online quiere segmentar a sus clientes para enviarles ofertas personalizadas según su perfil de viajero. Cuenta con información de 100 clientes. Utilizar el método jerárquico de Ward para agrupar a los clientes en segmentos, visualizar la estructura de agrupación con un dendrograma, y definir el número óptimo de clústeres.

1.0 Datos

```
import numpy as np import pan-
das as pd

import matplotlib.pyplot as plt import seaborn as
sns

from scipy.cluster.hierarchy import dendrogram, linkage, fcluster from sklearn.preprocessing import
StandardScaler

np.random.seed(42) n = 100

viajes = np.random.poisson(3, n).clip(min=1)

gasto = np.random.normal(1200, 300, n).clip(min=300) dias = np.random.normal(7,
2, n).clip(min=1)

tipo = np.random.choice([0, 1], size=n, p=[0.4, 0.6])

df_viajes = pd.DataFrame({ 'ViajesAnuales':viajes,
'GastoPromedio': gasto, 'DiasPromedio': dias, 'Tipo-
Viaje': tipo

})

df_viajes.head()
```

	ViajesAnuales	GastoPromedio	DiasPromedio	TipoViaje
0	4	1023.190573	4.359534	0
1	1	1454.880629	10.662918	1
2	3	1307.104646	9.358880	1

1.1 Estandarización

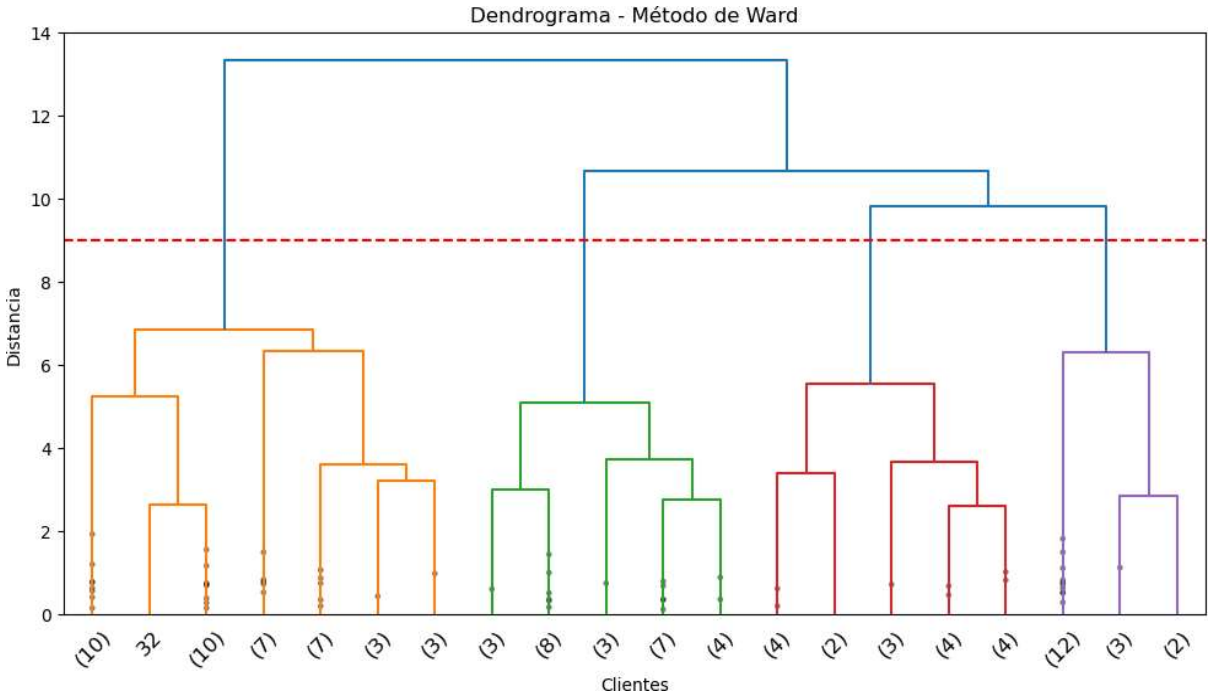
```
# Escalado

scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_viajes)
```

1.2 Ward

```
linked = linkage(X_scaled, method='ward')
```

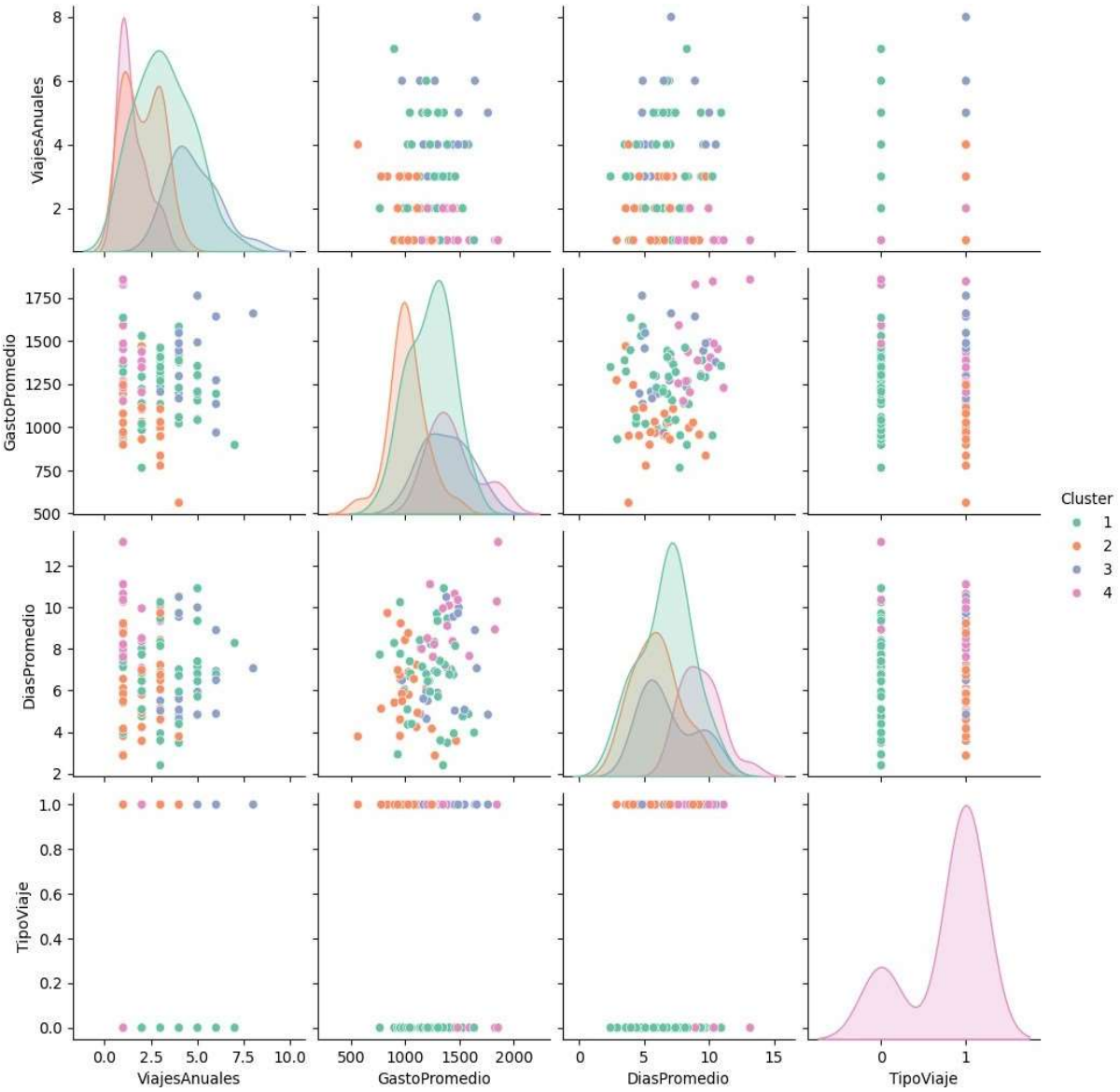
1.3 Dendrogrma

```
plt.figure(figsize=(12, 6))
dendrogram(linked, truncate_mode='lastp', p=20, leaf_rotation=45,
leaf_font_size=12, show_contracted=True)
plt.title("Dendrograma - Método de Ward")
plt.xlabel("Clientes") plt.ylabel("Distan-
cia")
plt.axhline(y=9, color='red', linestyle='--')  # línea de corte visual
plt.show()
```



1.4 Agrupamiento

```
# Elegir 4 clústeres
df_viajes['Cluster'] = fcluster(linked, t=4, criterion='maxclust')

# Ver agrupamiento
sns.pairplot(df_viajes, hue='Cluster', palette='Set2')
```



a) Número de clústeres 4 clústeres
b) Número de segmentos y su interpretación

- Viajes entre 2 y 7, Gasto entre 750 y 1500, Días entre 2 y 10 : Exploradores de escapadas [verde]
- Viajes entre 1 y 3, Gasto entre 500 y 1500, Días entre 2 y 10 : Turista eficiente [naranja]
- Viajes entre 2 y 9, Gasto entre 1000 y 2000, Días entre 4 y 12 : Aventureros premium [azul]
- Viajes entre 1 y 3, Gasto entre 1500 y 2000, Días entre 10 y 15 : Viajes de lujo [rosa]

Ejercicio de Análisis Multidimensional no métrico - Semejanza

Tenemos 4 productos: A, B, C y D.

Hicimos una encuesta a consumidores, y obtuvimos las siguientes **percepciones de disimilitud** entre ellos (en escala 1–9, donde 1 = muy parecidos, 9 = muy distintos):

	A	**B**	**C**	**D**
A	0	3	7	6
B	3	0	5	4
C	7	5	0	2
D	6	4	2	0

a) Usar **MDS no métrico** para representar perceptualmente los productos en 2D.
b) Visualizar el resultado con un mapa perceptual.
c) ¿Qué nos indica la cercanía entre dos productos en el mapa resultante?
d) ¿Qué producto parece estar más aislado? ¿Qué implicaciones podría tener para el marketing?
e) ¿Cómo podrías usar esta información para segmentar el mercado o diseñar campañas?

Ejercicio Resuelto

Tenemos 4 productos: A, B, C y D. Hicimos una encuesta a consumidores, y obtuvimos las siguientes percepciones de disimilitud entre ellos (en escala 1–9, donde 1 = muy parecidos, 9 = muy distintos)

1.0 Datos

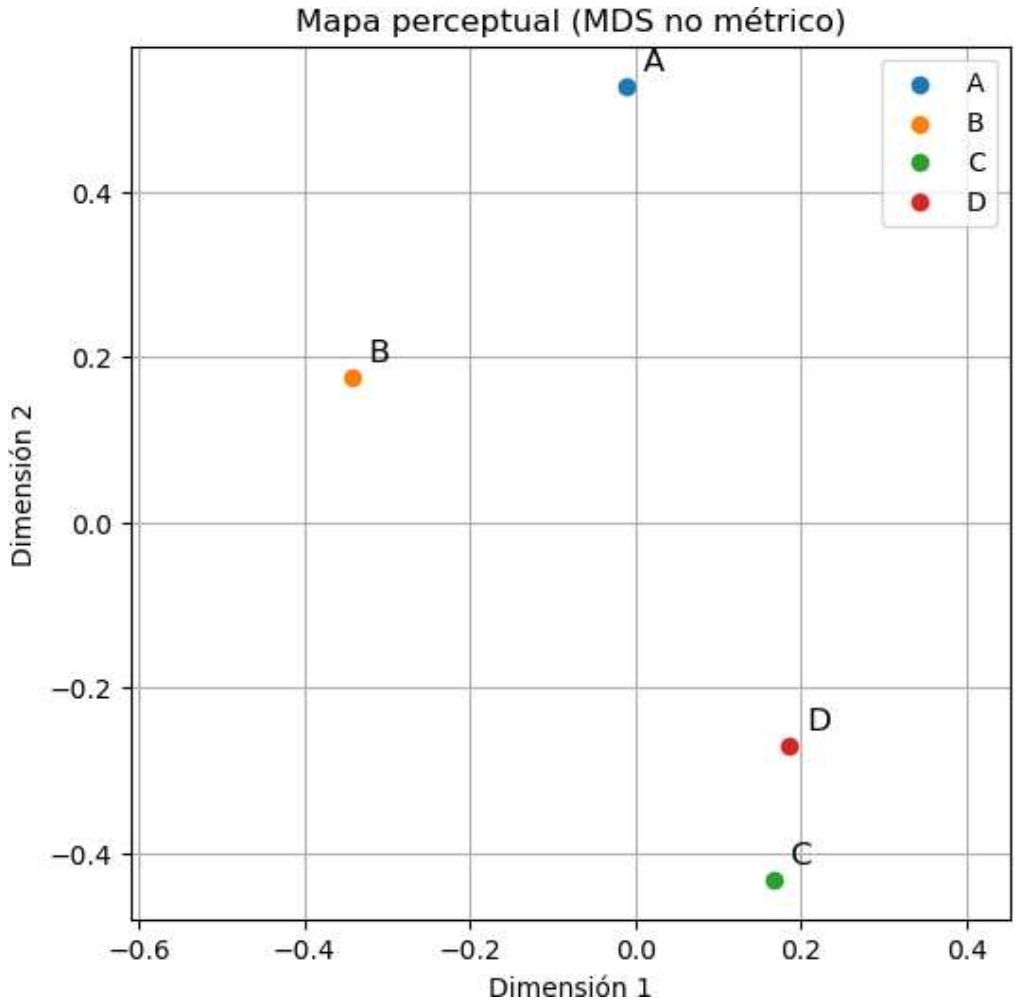
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.manifold import MDS

# Paso 1: Matriz de disimilitudes

dissimilarities = np.array([
[0, 3, 7, 6],
[3, 0, 5, 4],
[7, 5, 0, 2],
[6, 4, 2, 0]
])

labels = ['A', 'B', 'C', 'D']
```

1.1 MDS no métrico

```
mds = MDS(n_components=2, dissimilarity='precomputed', metric=False,
random_state=42)
coords = mds.fit_transform(dissimilarities)
```

1.2 Visualización

```
plt.figure(figsize=(6, 6))
for i, label in enumerate(labels):
x, y = coords[i] plt.scatter(x, y,
label=label)
plt.text(x + 0.02, y + 0.02, label, fontsize=12)

plt.title("Mapa perceptual (MDS no métrico)")
plt.xlabel("Dimensión 1")
plt.ylabel("Dimensión 2")
plt.grid(True) plt.leg-
end() plt.axis('equal')
plt.show()
```



a) Usar MDS no métrico para representar perceptualmente los productos en 2D.

b) Visualizar el resultado con un mapa perceptual.

c) ¿Qué nos indica la cercanía entre dos productos en el mapa resultante?

C y D están cerca en el mapa perceptual. Esto nos indica que ambos productos son muy parecidos en la mente de los consumidores. Esta cercanía nos dice que podemos usar la misma estrategia de marketing para ambos o, incluso, combinarlos en una única campaña.

d) ¿Qué producto parece estar más aislado? ¿Qué implicaciones podría tener para el marketing?

El más aislado parece ser A. Es diametralmente opuesto a C y D, por lo que podría indicar que aquellos consumidores que compran C y D no estén interesados en A. Así mismo, también podríamos tener un nicho de mercado no cubierto y, por ello, A queda fuera de la imagen mental que los consumidores tienen de la marca.

Como el producto B está cerca de A, podemos asumir que A satisface necesidades no atendidas por C y D, por lo que conviene estudiar sus diferencias. Esto podría darnos la opción de crear un nuevo producto situado en la mente de los consumidores entre A y C y D.

Así mismo, del producto B extraemos un análisis similar al estudiar la dimensión 1.

e) ¿Cómo podrías usar esta información para segmentar el mercado o diseñar campañas?

- Creando una campaña conjunta para los productos C y D

- Se podrían analizar las necesidades que cubren en los consumidores cada uno de los productos y crear, como puntos intermedios, nuevos productos que atraigan a nuevos nichos de mercado.
- Segmentar los consumidores entre: compradores de C y D, compradores de B y compradores de A.

Ejercicio de Análisis Multidimensional no métrico - Preferencia

3 consumidores califican 4 productos (A, B, C, D) en una escala de 1 a 10 (mayor valor = mayor preferencia):

Producto	**Persona 1**	**Persona 2**	**Persona 3**
A	8	7	6
B	6	5	4
C	4	8	7
D	2	3	5

Preguntas

a) Usar **MDS no métrico** para representar perceptualmente los productos en 2D.
b) Visualizar el resultado con un mapa perceptual.
c) ¿Qué nos indica la cercanía entre dos productos en el mapa resultante?
d) ¿Qué producto parece estar más aislado? ¿Qué implicaciones podría tener para el marketing?
e) ¿Cómo podrías usar esta información para segmentar el mercado o diseñar campañas?

Ejercicio Resuelto

3 consumidores califican 4 productos (A, B, C, D) en una escala de 1 a 10 (mayor valor = mayor preferencia)

1.0 Datos

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import pairwise_distances
from sklearn.manifold import MDS

# Datos de preferencias (filas = productos, columnas = consumidores)

preferences = np.array([
[8, 7, 6], # A
[6, 5, 4], # B

[4, 8, 7], # C

[2, 3, 5], # D
```

1.1 Matriz de disimilitud

```
dissimilarities = pairwise_distances(preferences, metric='euclidean')
```

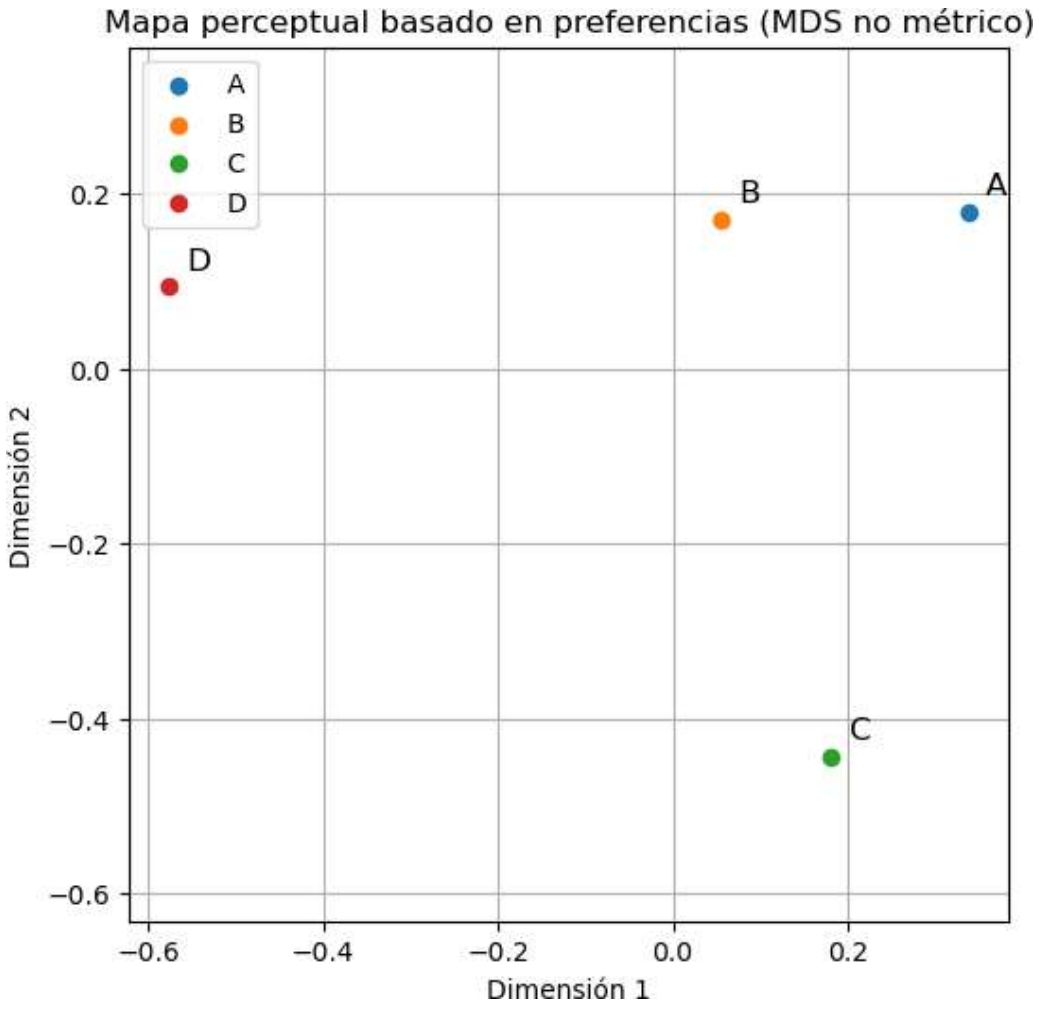
1.2 MDS

```
mds = MDS(n_components=2, dissimilarity='precomputed', metric=False,
random_state=42)
coords = mds.fit_transform(dissimilarities)
```

1.3 Visualización

```
plt.figure(figsize=(6, 6))
for i, label in enumerate(labels):
x, y = coords[i] plt.scatter(x, y,
label=label)
plt.text(x + 0.02, y + 0.02, label, fontsize=12)

plt.title("Mapa perceptual basado en preferencias (MDS no métrico)")
plt.xlabel("Dimensión 1")
plt.ylabel("Dimensión 2")
```

```
plt.grid(True) plt.leg-
end() plt.axis('equal')
plt.show()
```



a) Usar MDS no métrico para representar perceptualmente los productos en 2D.

b) Visualizar el resultado con un mapa perceptual.

c) ¿Qué nos indica la cercanía entre dos productos en el mapa resultante?

Los dos productos más cercanos son A y B. Esto podría indicarnos que están satisfaciento la misma necesidad en la mente de los consumidores.

d) ¿Qué producto parece estar más aislado? ¿Qué implicaciones podría tener para el marketing? Observando las dimensiones por separado:

- En la dimensión 1 se separa claramente el producto D de los otros tres
- En la dimensión 2 se separa claramente el producto C de los otros tres

Por lo que podemos deducir que los productos A y B hacen de conexión entre ellos.

Que haya productos aislados puede implicar que son nichos del mercado a los que podemos acceder o, por el

contrario, que estos produtos no están bien cimentados en la mente de los consumidores.

En este caso, como los productos A y B hacen de conexión, nos inclinamos más hacia la idea de que son nichos de mercado a los que podemos acceder y con los que podríamos incrementar el alcance de la empresa.

e) ¿Cómo podrías usar esta información para segmentar el mercado o diseñar campañas?

- Hacer campañas que unifiquen los productos A, B y C
- Hacer campañas que unifiquen los productos A, B y D
- Crear nuevos productos con característica que puedan satisfacer a los compradores de C y D

Ejercicio de CHAID

Una tienda desea comprender qué características de sus clientes influyen en la decisión de comprar un producto específico. Para ello, ha recolectado datos categóricos de clientes incluyendo **edad**, **nivel de ingreso**, **género**, y **frecuencia de compra**. Se desea construir un modelo de árbol de decisión que permita **segmentar a los clientes** según estas variables y así identificar patrones que expliquen la compra del producto.

Edad	**Ingreso**	**Género**	**FrecuenciaCompra**	**CompraProducto**
<30	Bajo	Mujer	Alta	Sí
30-45	Medio	Hombre	Baja	No
30-45	Medio	Mujer	Media	Sí
>45	Alto	Hombre	Alta	Sí
<30	Bajo	Mujer	Media	No
>45	Medio	Mujer	Baja	No
30-45	Alto	Hombre	Alta	Sí
<30	Bajo	Hombre	Media	No
>45	Medio	Mujer	Baja	No
30-45	Alto	Mujer	Alta	Sí

Preguntas

a) Construye un árbol CHAID usando como objetivo la respuesta a la campaña anterior.
b) Interpreta los segmentos:
    a. ¿Qué combinaciones de variables están asociadas a una alta tasa de compra?
    b. ¿Qué segmentos tienen muy baja compra?
c) Recomienda una estrategia de marketing basada en los grupos de mayor probabilidad de compra.

Ejercicio Resuelto

Una tienda desea comprender qué características de sus clientes influyen en la decisión de comprar un producto específico. Para ello, ha recolectado datos categóricos de clientes incluyendo edad, nivel de ingreso, género, y frecuencia de compra. Se desea construir un modelo de árbol de decisión que permita segmentar a los clientes según estas variables y así identificar patrones que expliquen la compra del producto.

1.0 Datos

```
import pandas as pd

from sklearn.tree import DecisionTreeClassifier, export_text, export_graphviz

from sklearn.preprocessing import LabelEncoder import matplotlib.pyplot
as plt

from sklearn.tree import plot_tree

# Datos con variables ya categorizadas

data = pd.DataFrame({

'Edad': ['<30', '30-45', '30-45', '>45', '<30', '>45', '30-45',

'<30', '>45', '30-45'],

'Ingreso': ['Bajo', 'Medio', 'Medio', 'Alto', 'Bajo', 'Medio', 'Alto', 'Bajo', 'Medio', 'Alto'],

'Genero': ['Mujer', 'Hombre', 'Mujer', 'Hombre', 'Mujer', 'Mujer', 'Hombre', 'Hombre', 'Mujer', 'Mujer'],

'FrecuenciaCompra': ['Alta', 'Baja', 'Media', 'Alta', 'Media',

'Baja', 'Alta', 'Media', 'Baja', 'Alta'],

'CompraProducto': ['Sí', 'No', 'Sí', 'Sí', 'No', 'No', 'Sí', 'No', 'No', 'Sí']

})Edad Ingreso  Genero FrecuenciaCompra CompraProducto

df0      <30       Bajo       Mujer                Alta              Sí
 1    30-45      Medio   Hombre                 Baja              No
 2    30-45      Medio       Mujer              Media              Sí
 3      >45       Alto   Hombre                 Alta              Sí
 4      <30       Bajo       Mujer              Media              No
 5      >45      Medio       Mujer                Baja              No
 6    30-45       Alto   Hombre                 Alta              Sí
 7      <30       Bajo   Hombre                 Media              No
 8      >45      Medio       Mujer                Baja              No
```

1.1 Preparar datos

```
# Mapeo manual para 'FrecuenciaCompra' mapa_frecuencia = {'Baja': 0, 'Media': 1,
'Alta': 2} data['FrecuenciaCompra'] = data['FrecuenciaCompra'].map(mapa_frecuencia)

# Codificar el resto de columnas con LabelEncoder excepto 'FrecuenciaCompra'
label_encoders = {}
for col in data.columns:
if col != 'FrecuenciaCompra': le = LabelEncoder()
data[col] = le.fit_transform(data[col]) label_encoders[col] = le  # Guardar encoder para futura
decodificación

# Confirmar qué valor representa cada clase en la variable objetivo clases_objetivo = label_encoders['Com-
praProducto'].classes_ print("Clases originales de 'CompraProducto':")
print(f"0 = {clases_objetivo[0]}") print(f"1 = {clases_obje-
tivo[1]}")

# Separar variables independientes y objetivo
X = data.drop('CompraProducto', axis=1) y = data['CompraProduc-
to']

print(X) print('\n', y)
```

Clases originales de 'CompraProducto':
0 = No

	Edad	Ingreso	Genero	FrecuenciaCompra
0	1	1	1	2
1	0	2	0	0
2	0	2	1	1
3	2	0	0	2
4	1	1	1	1
5	2	2	1	0

0	1
1	0
2	1
3	1
4	0
5	0

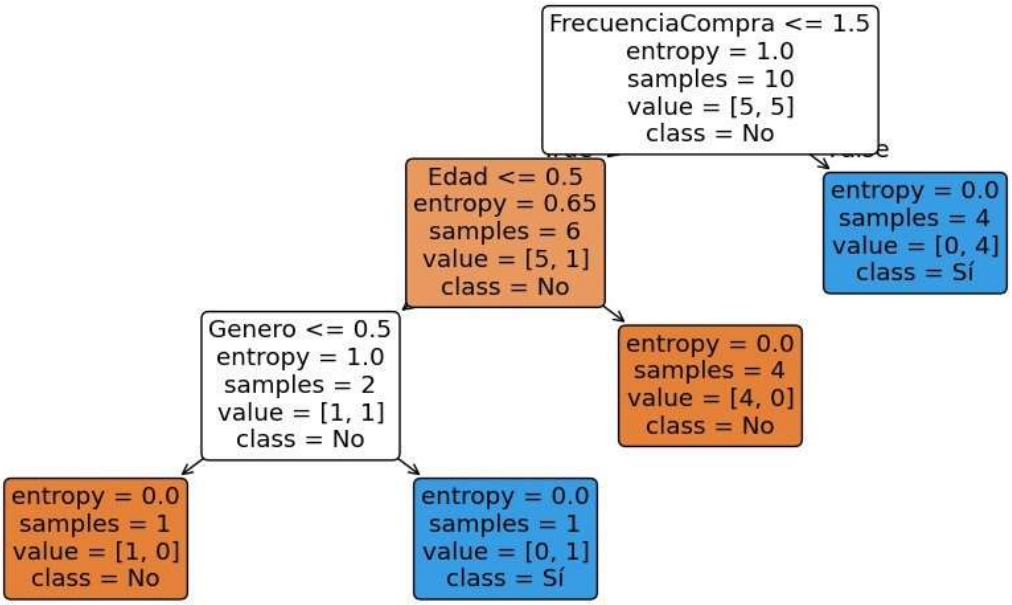
```
6      1
7      0
8      0
9      1
Name: CompraProducto, dtype: int64
```

1.2 Árbol

```
# 4. Crear y entrenar el modelo

clf = DecisionTreeClassifier(criterion='entropy', max_depth=3, random_state=0)

clf.fit(X, y)

# 5. Mostrar árbol en texto

print(export_text(clf, feature_names=list(X.columns)))

|--- FrecuenciaCompra <= 1.50

|       |--- Edad <= 0.50

|       |       |--- Genero <= 0.50

|       |       |       |--- class: 0

|       |       |--- Genero >  0.50

|       |       |       |--- class: 1
```

1.3 Árbol

```
plt.figure(figsize=(12, 6))
plot_tree(clf,
feature_names=X.columns, class_names=label_encoders['CompraProduc-
to'].classes_, filled=True,
rounded=True)
plt.show()
```



a) Construye un árbol CHAID usando como objetivo la respuesta a la campaña anterior

b) Interpreta los segmentos:

a. ¿Qué combinaciones de variables están asociadas a una alta tasa de compra?

Cualquier combinación con Frecuencia de compra > 1.5

b. ¿Qué segmentos tienen muy baja compra?

Cualquier combinación con Frecuencia de compra <= 1.5

c) Recomienda una estrategia de marketing basada en los grupos de mayor probabilidad de compra.

Aumentar la promoción y sacar nuevos productos, ya que son clientes acostumbrados a comprar, probablemente estarán atraídos por las novedades.

Nota: Este cógido se ha realizado con DecisionTreeClassifier

Este árbol de decisión se parece al que genera CHAID porque ambos métodos construyen una estructura jerárquica que divide los datos en segmentos homogéneos, basándose en variables predictoras categóricas. Aunque CHAID utiliza pruebas estadísticas de chi-cuadrado para determinar las divisiones más significativas y puede generar nodos con más de dos ramas (multivía), mientras que el árbol de decisión de scikit-learn por defecto genera divisiones binarias, ambos modelos buscan maximizar la pureza o separación entre clases en cada nivel del árbol. Visualmente y conceptualmente, los árboles resultantes representan reglas de decisión claras para segmentar clientes, lo cual es esencial en marketing. Por eso, incluso con métodos distintos, el árbol generado en Python puede cumplir una función similar a un árbol CHAID clásico.

Anexo III.
Examen

Ejercicio 1

Se quiso estudiar la relación entre edad del cliente y canal preferido de compra en una papelería obteniendo la siguiente tabla de contingencia:

Canal de compra / Edad	**Tienda física**	**Página web**	**App móvil**	**Redes sociales**	**Total**
18–25 años	10	15	25	20	70
26–35 años	20	30	35	10	95
36–50 años	25	35	15	5	80
51+ años	30	20	5	0	55
Total	85	100	80	35	300

a) Aplicando la técnica de tabulación cruzada, ¿a qué segmentos recomendarías prestar más atención? Incluye al menos un gráfico.
b) Comprueba si vale la pena hacer dos campañas de promoción diferenciadas. ¿Cuáles serían?

Nota: El valor de la chi cuadrado al 95% de confianza con 2 grados de libertad es 5,99; y con 3 grados de libertad es 7,81; con 9 grados de libertad es de 16,92. Considere trabajar con 2 cifras decimales.

Ejercicio 2

Supongamos que tenemos una tabla de contingencia donde las filas representan canales de venta y las columnas, tipos de productos.

	Tienda Física	Web	App
Producto A	30	5	10
Producto B	10	15	5
Producto C	5	20	25
Producto D	5	10	15

Preguntas

a) ¿Qué productos están más asociados con ciertos canales de venta?

b) ¿Se parecen los patrones de venta entre Web y App?

c) ¿Hay productos que se distribuyan por todos los canales por igual?

d) ¿Algún producto depende fuertemente de un solo canal?

e) ¿Cómo se visualiza gráficamente la afinidad entre un producto y un canal?

f) ¿Qué representa cada componente?

Ejercicio 3

Una plataforma de cursos online quiere segmentar a sus usuarios para personalizar recomendaciones y ofertas de suscripciones. Se tiene información sobre el comportamiento de 120 usuarios.

Variables disponibles:

- Cursos completados en los últimos 6 meses
- Horas promedio dedicadas por semana
- Tasa de finalización de cursos (en %)
- Suscripción activa (0 = No, 1 = Sí)

Aquí se puede leer un extracto:

CursosCompletados	HorasSemana	TasaFinalizacion	SuscripcionActiva
7	9.909299	45.179528	1
5	6.347104	86.515438	1
4	6.066840	67.212536	0
8	11.994263	67.496876	1

Aplicar el método jerárquico de **Ward** para identificar perfiles como: estudiantes activos, suscriptores inactivos, usuarios casuales, etc.

Ejercicio 4

Una empresa de calzado deportivo quiere entender cómo se posicionan diferentes **tipos de zapatillas** en la mente de los consumidores en función de sus **preferencias generales** (gustos). Para ello, ha recolectado datos sobre preferencias de un grupo de consumidores hacia distintos tipos de zapatillas, y a partir de esos datos ha construido una **matriz de disimilitudes**, donde los valores indican **qué tan diferentes son dos tipos de zapatillas en términos de su atractivo para los consumidores**.

Matriz de disimilitudes (ya procesada):

	Running	**Urbanas**	**Senderismo**	**Skate**	**Minimalistas**
Running	0	2.0	4.0	3.5	2.5
Urbanas	2.0	0	3.0	1.5	1.0
Senderismo	4.0	3.0	0	3.5	4.0
Skate	3.5	1.5	3.5	0	2.0
Minimalistas	2.5	1.0	4.0	2.0	0

Construye un **mapa perceptual** usando **MDS no métrico** para visualizar cómo se diferencian estos tipos de zapatillas **según las preferencias** de los consumidores y explica todo lo que veas en él.

Ejercicio 1

2.0 Datos

```
# Importar librerías necesarias

import pandas as pd im-
port numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency

# Datos: Edad vs Canal de compra preferido

data_edad_canal = {
'Edad': ['18-25'] * 4 + ['26-35'] * 4 + ['36-50'] * 4 + ['51+'] *
4,
'Canal': ['Tienda', 'Web', 'App', 'RRSS'] * 4,
'Frecuencia': [10, 15, 25, 20, 20, 30, 35, 10, 25, 35, 15, 5, 30,
20, 5, 0]
}
```

2.1 Frecuencias porcentuales

```
# Tabla de frecuencias absolutas

tabla_abs = pd.pivot_table(df, values='Frecuencia', index='Edad', co-
lumns='Canal', aggfunc=np.sum, fill_value=0)

# Porcentajes por fila

tabla_fila = tabla_abs.div(tabla_abs.sum(axis=1), axis=0) * 100

# Porcentajes por fila

tabla_columna = tabla_abs.div(tabla_abs.sum(axis=0), axis=1) * 100

# Porcentajes globales

tabla_global = (tabla_abs / tabla_abs.values.sum()) * 100

C:\Users\San\AppData\Local\Temp\ipykernel_248\645069978.py:2: Future-
Warning: The provided callable <function sum at 0x0000018D0C2FCFE0> is
currently using DataFrameGroupBy.sum. In a future version of pandas,
the provided callable will be used directly. To keep current behavior
```

2.2 Test Chi-cuadrado

```
# Prueba chi-cuadrado

chi2, p, dof, expected = chi2_contingency(tabla_abs)
```

2.3 Resultados

```
# Mostrar resultados
print("Tabla de frecuencias absolutas:\n", tabla_abs, "\n") print("Porcentajes por fila (%):\n", tabla_fila.round(1),
"\n") print("Porcentajes por columna (%):\n", tabla_columna.round(1), "\n") print("Porcentajes globales (%):\n",
tabla_global.round(1), "\n") print(f"Chi-cuadrado = {chi2:.2f}, p-valor = {p:.4f}, grados de libertad = {dof}",
"\n")
print("Frecuencias esperadas bajo H0:\n", pd.DataFrame(expected, index=tabla_abs.index, columns=tabla_abs.co-
lumns).round(2))
```

Tabla de frecuencias absolutas:

Canal	App	RRSS	Tienda	Web	Edad
18-25	2	20	10	1	
	5			5	
26-35	3	10	20	3	
	5			0	
36-50	1	5	25	3	
	5			5	
51+	5	0	30	2	
				0	

Porcentajes por fila (%):

Canal	App	RRSS	Tienda	Web	Edad
18-25	35.7	28.6	14.3	21.4	
26-35	36.8	10.5	21.1	31.6	
36-50	18.8	6.2	31.2	43.8	
51+	9.1	0.0	54.5	36.4	

Porcentajes por columna (%):

Canal	App	RRSS	Tienda	Web	Edad
18-25	31.2	57.1	11.8	15.0	
26-35	43.8	28.6	23.5	30.0	
36-50	18.8	14.3	29.4	35.0	
51+	6.2	0.0	35.3	20.0	

Porcentajes globales (%):

Canal	App	RRSS	Tienda	Web	Edad
18-25	8.3	6.7	3.3	5.0	
26-35	11.7	3.3	6.7	10.0	
36-50	5.0	1.7	8.3	11.7	
51+	1.7	0.0	10.0	6.7	

Chi-cuadrado = 65.82, p-valor = 0.0000, grados de libertad = 9

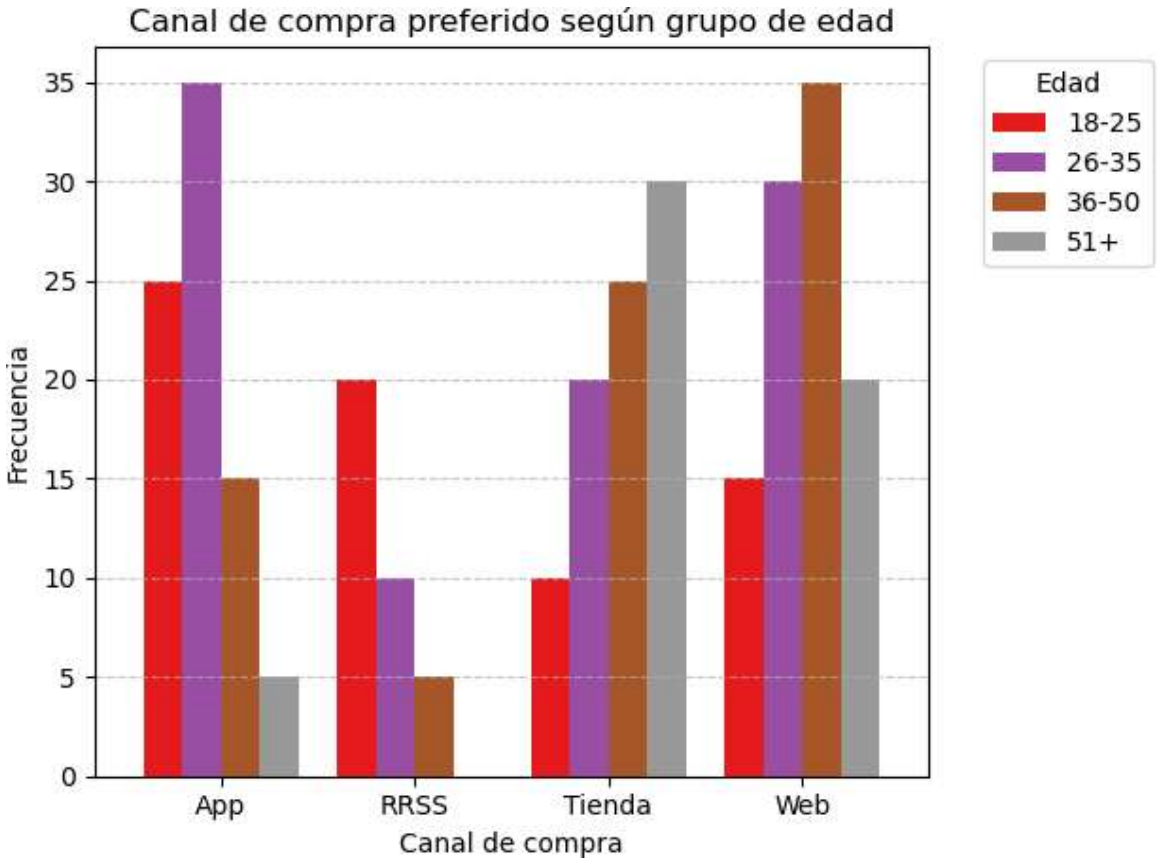
Frecuencias esperadas bajo H0:

Canal	App	RRSS	Tienda	Web
18-25	18.67	8.17	19.83	23.33
26-35	25.33	11.08	26.92	31.67
36-50	21.33	9.33	22.67	26.67
51+	14.67	6.42	15.58	18.33

2.4 Gráficos

```
# Gráfico de barras no apiladas

plt.figure(figsize=(10, 6))
tabla_abs.T.plot(kind='bar', colormap='Set1', width=0.8) plt.title('Ca-
nal de compra preferido según grupo de edad') plt.ylabel('Frecuencia')
plt.xlabel('Canal de compra') plt.xticks(rota-
tion=0)
plt.legend(title='Edad', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```


Canal de compra preferido según grupo de edad
Edad
18-25
26-35
36-50
51+
Frecuencia
0
5
10
15
20
25
30
35
App
RRSS
Tienda
Web
Canal de compra

Ejercicio 2

2.0 Datos

```
import pandas as pd import numpy
as np

import matplotlib.pyplot as plt import seaborn as sns

from sklearn.preprocessing import StandardScaler

import prince  # Librería para Análisis de Correspondencias

# Tabla de contingencia

data = {

'Producto A': [30, 5, 10],

'Producto B': [10, 15, 5],
```

```
print(df)
```

	Producto A	Producto B	Producto C	Producto D
Tienda Física	30	10	5	5
Web	5	15	20	10
App	10	5	25	15

2.1 Análisis de Correspondencias

```
# Crear el modelo MCA con sklearn como motor

mca = prince.MCA( n_com-
ponents=2, en-
gine='sklearn', ran-
dom_state=42
)

# Ajustar el modelo

mca = mca.fit(df)

# Obtener coordenadas de individuos (filas)

row_coords = mca.row_coordinates(df)

# Obtener coordenadas de categorías (columnas)
```

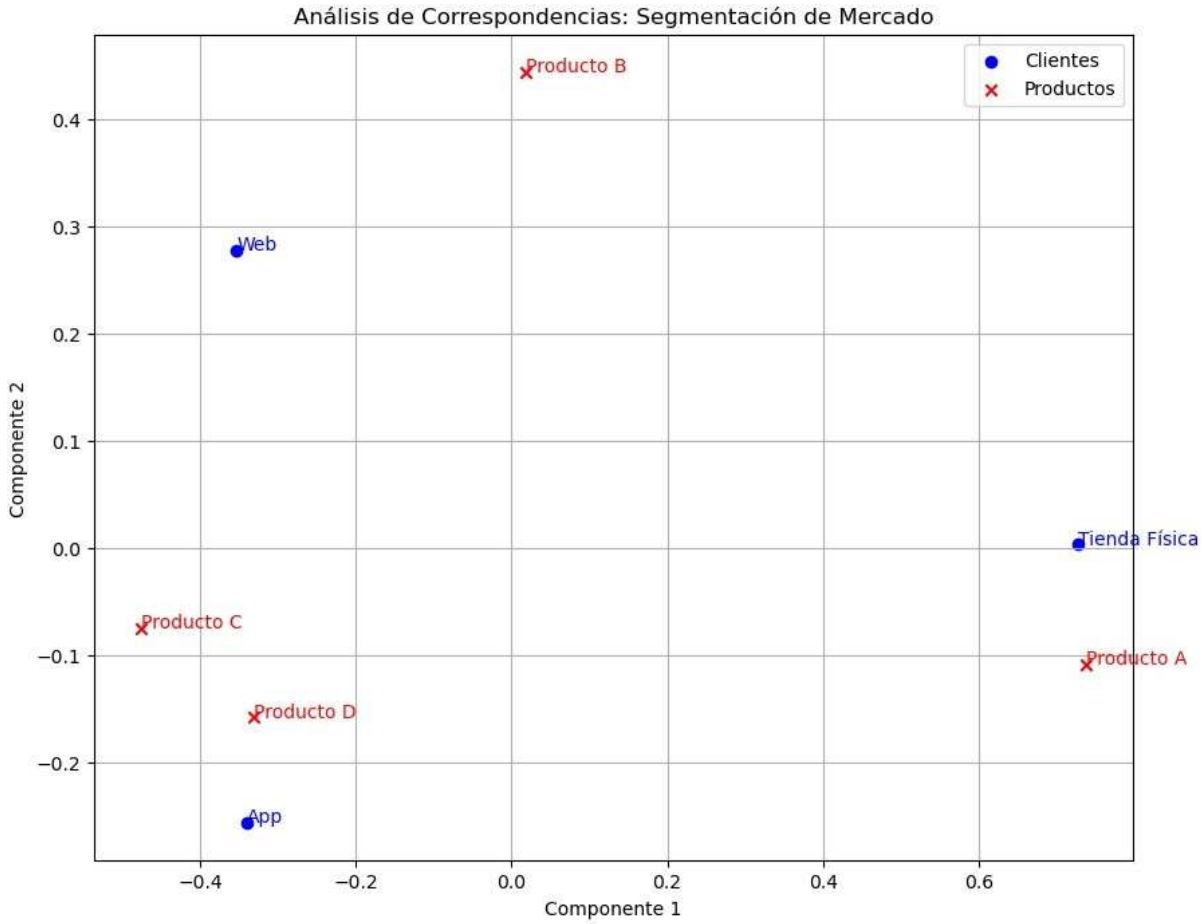
```
print("Coordenadas de los modos de venta:\n", row_coords) print("\nCoordenadas de los productos:\n", col_coords)
```

```
                Coordenadas de los modos de venta:
                                        0         1
Tienda Física   0.727545   0.003328
```

```
Coordenadas de los productos:
0                                       1
Producto A   0.737241 -0.108141
Producto B  0.018357  0.444286
```

2.2 Gráfica

```
# Graficar resultados
plt.figure(figsize=(10, 8))

# Filas
plt.scatter(row_coords[0], row_coords[1], c='blue', label='Clientes')
for i, txt in enumerate(df.index):
plt.annotate(txt, (row_coords.iloc[i, 0], row_coords.iloc[i, 1]),
color='blue')

# Columnas
plt.scatter(col_coords[0], col_coords[1], c='red', label='Productos',
marker='x')
for i, txt in enumerate(df.columns):
plt.annotate(txt, (col_coords.iloc[i, 0], col_coords.iloc[i, 1]),
color='red')

plt.title('Análisis de Correspondencias: Segmentación de Mercado')
plt.xlabel('Componente 1')
plt.ylabel('Componente 2')
plt.legend()
```



2.3 Contribuciones

```
col_coords = mca.column_coordinates(df)

# Cuadrados de las coordenadas

col_coords_sq = col_coords**2

# Contribución de cada categoría a cada dimensión (en %) dim1_contrib = 100 * col_coords_sq[0] /
col_coords_sq[0].sum() dim2_contrib = 100 * col_coords_sq[1] / col_coords_sq[1].sum()

# Mostrar como tabla en vez de usar print con formato de número

contrib_df = pd.DataFrame({
'Contribución Dim 1 (%)': dim1_contrib, 'Contribu-
ción Dim 2 (%)': dim2_contrib
})

print(contrib_df.round(2))

# Varianza de cada dimension total_var_dim1 =
sum(col_coords[0]**2) total_var_dim2 =
```

```
print(f"\nVarianza aproximada Dim 1: {total_var_dim1:.4f}")\nprint(f"\nVarianza aproxi-
```

	Contribución Dim 1 (%)	Contribución	Dim	2 (%)
Producto A	61.76			4.88
Producto B	0.04			82.39
Producto C	25.68			2.32
Producto D	12.52			10.41

```
Varianza aproximada Dim 1: 0.8800
```

Ejercicio 3

2.0 Datos

```
import numpy as np import pandas
as pd

import matplotlib.pyplot as plt import seaborn as
sns

from sklearn.preprocessing import StandardScaler

from scipy.cluster.hierarchy import linkage, dendrogram, fcluster


np.random.seed(123) n = 120

cursos = np.random.poisson(5, n).clip(min=0) horas_semana = np.random.normal(6,
2.5, n).clip(min=0.5)

tasa_finalizacion = np.random.normal(75, 15, n).clip(0, 100) suscripcion = np.random.choice([0, 1], size=n,
p=[0.4, 0.6])

df_edu = pd.DataFrame({ 'CursosCompletados': cursos, 'Horas-
Semana': horas_semana, 'TasaFinalizacion': tasa_finalizacion,
'SuscripcionActiva': suscripcion

})

df_edu.head()
```

	CursosCompletados	HorasSemana	TasaFinalizacion	SuscripcionActiva
0	7	9.909299	45.179528	1
1	5	6.347104	86.515438	1
2	4	6.066840	67.212536	0

2.1 Estandarización

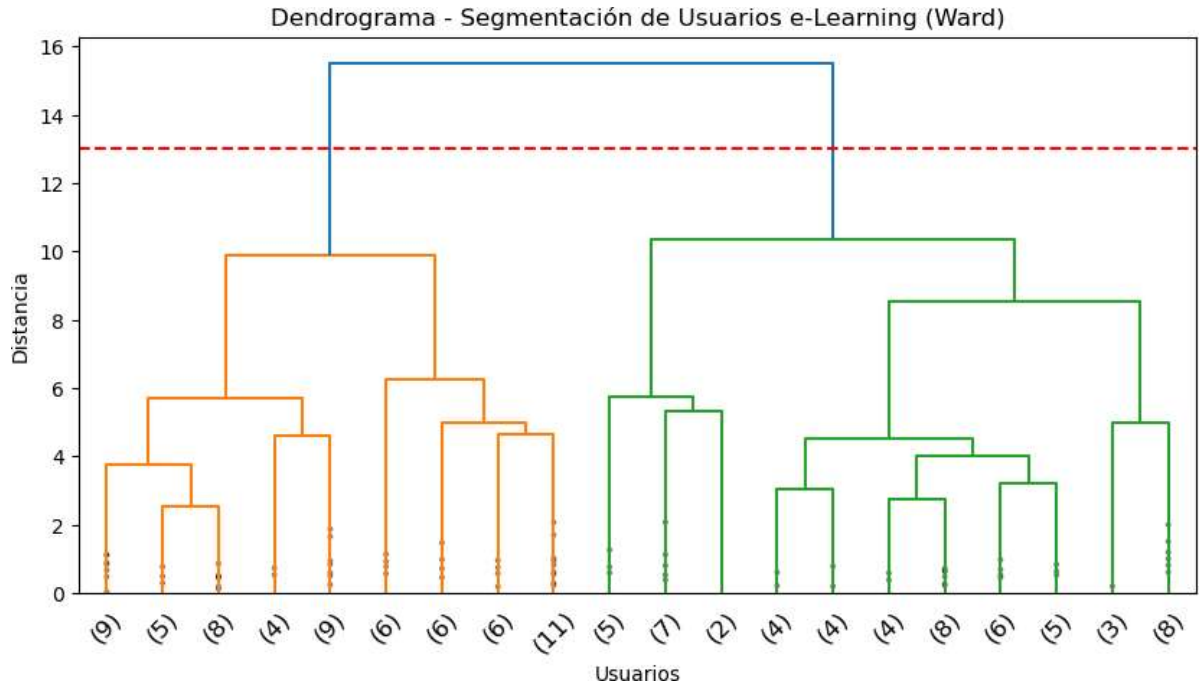
```
# Escalado

scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_edu)
```

2.2 Ward

```
linked = linkage(X_scaled, method='ward')
```

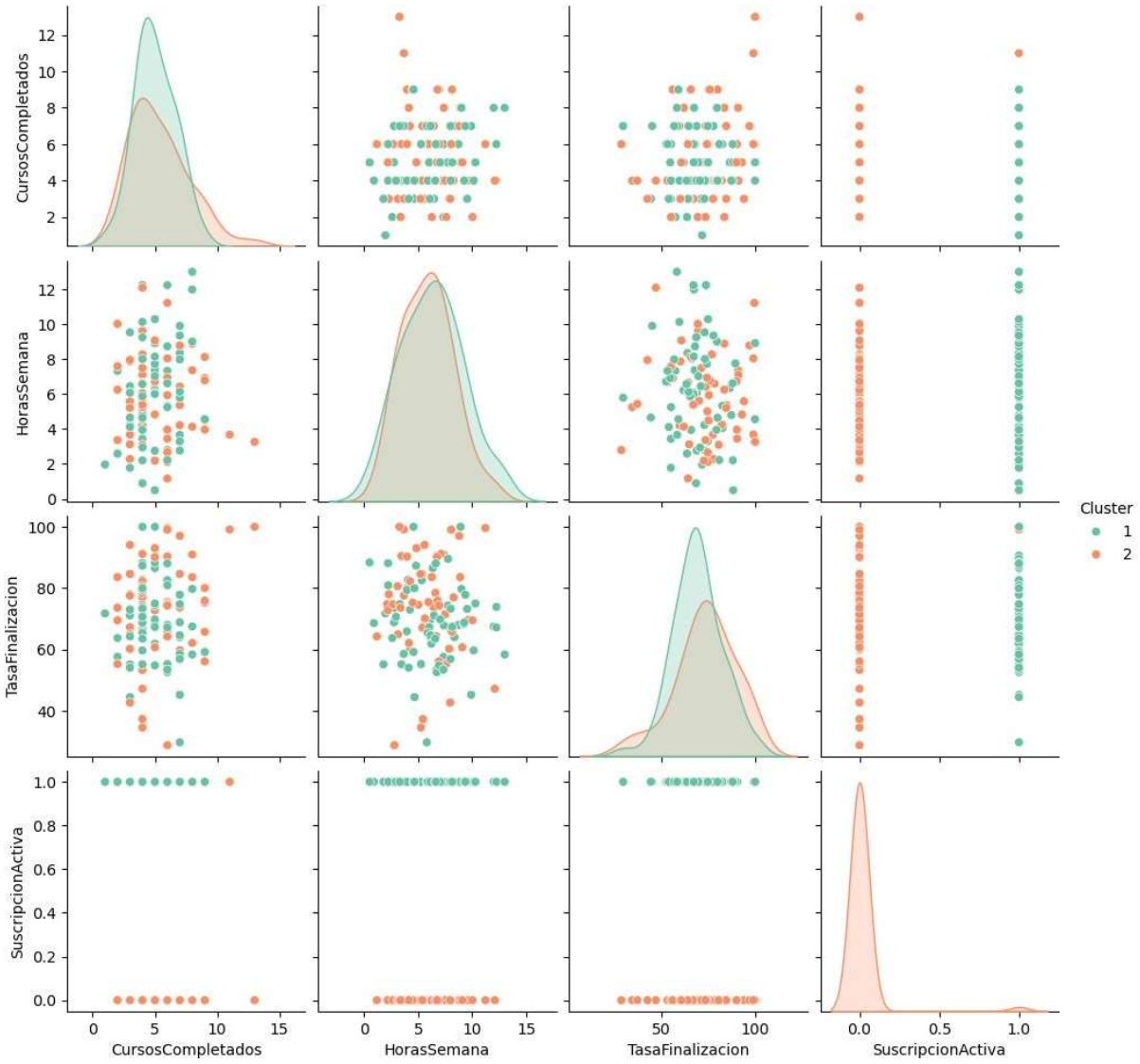
2.3 Dendrogrma

```
plt.figure(figsize=(10, 5))
dendrogram(linked, truncate_mode='lastp', p=20, leaf_rotation=45,
show_contracted=True)
plt.axhline(y=13, color='red', linestyle='--') plt.title("Dendrograma
- Segmentación de Usuarios e-Learning (Ward)") plt.xlabel("Usuarios")
plt.ylabel("Distancia")
plt.show()
```



2.4 Agrupamiento

```
df_edu['Cluster'] = fcluster(linked, t=2, criterion='maxclust')

# Ver agrupamiento

sns.pairplot(df_edu, hue='Cluster', palette='Set2')
plt.show()
```


CursosCompletados
HorasSemana
TasaFinalizacion
SuscripcionActiva
Cluster
1
2

Ejercicio 4

2.0 Datos

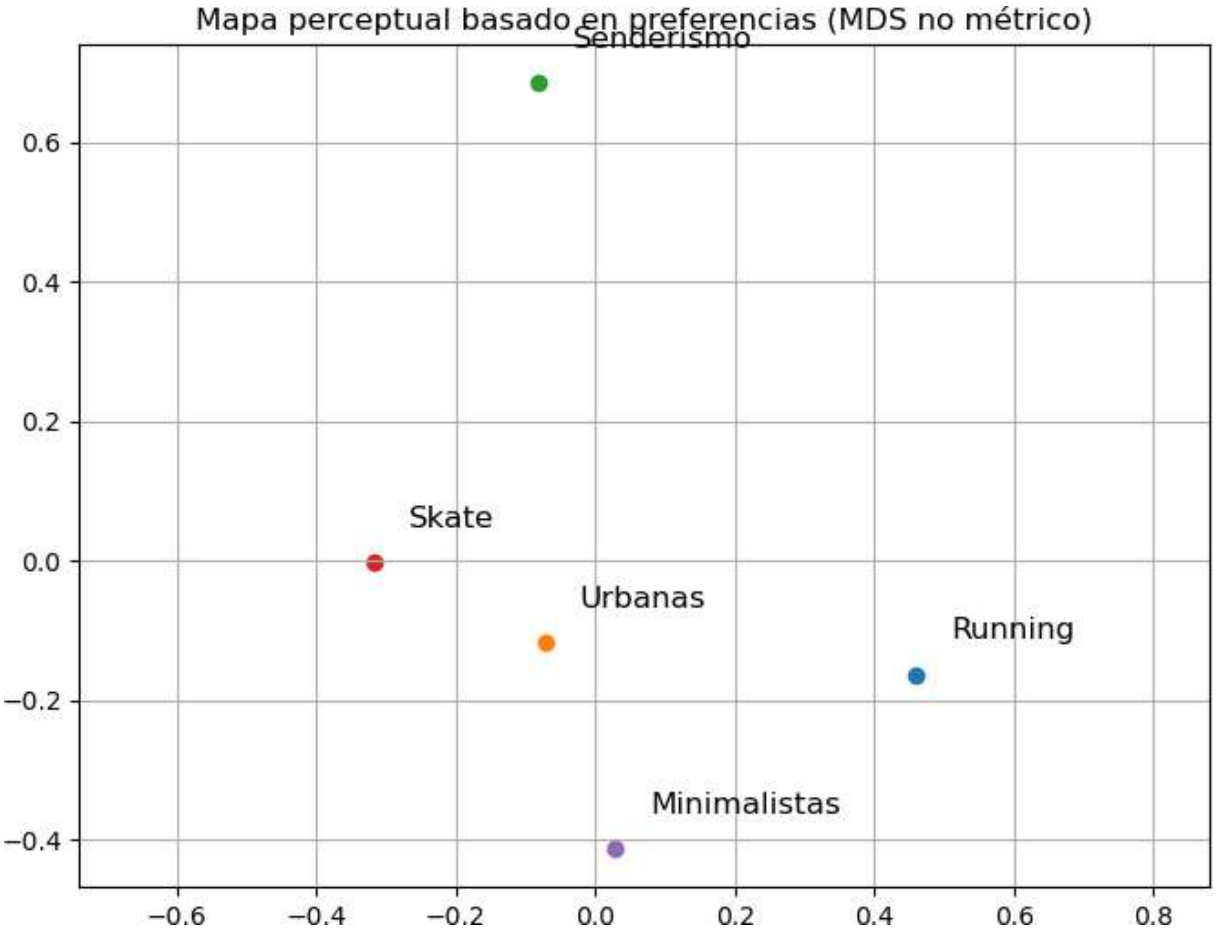
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.manifold import MDS

# Paso 1: Matriz de disimilitudes basada en preferencias

dissimilarities = np.array([
[0.0, 2.0, 4.0, 3.5, 2.5],
[2.0, 0.0, 3.0, 1.5, 1.0],
[4.0, 3.0, 0.0, 3.5, 4.0],
[3.5, 1.5, 3.5, 0.0, 2.0],
[2.5, 1.0, 4.0, 2.0, 0.0]
])

labels = ['Running', 'Urbanas', 'Senderismo', 'Skate', 'Minimalistas']

# Paso 2: Aplicar MDS no métrico # Paso 3: Visua-

lización
```

2.1 MDS

```
mds = MDS(n_components=2, dissimilarity='precomputed', metric=False,
random_state=42)
coords = mds.fit_transform(dissimilarities)
```

2.2 Visualización

```
plt.figure(figsize=(8, 6))
for i, label in enumerate(labels):
x, y = coords[i] plt.scatter(x, y)
plt.text(x + 0.05, y + 0.05, label, fontsize=12)

plt.title("Mapa perceptual basado en preferencias (MDS no métrico)")
plt.grid(True)
plt.axis('equal')
plt.show()
```


Mapa perceptual basado en preferencias (MDS no métrico)
Senderismo
Skate
Urbanas
Running
Minimalistas
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Anexo IV.
Especificaciones técnicas de la plataforma

En este documento se detallan los costes tecnológicos asociados al desarrollo del proyecto, así como una explicación de los sistemas utilizados. Además, permite justificar las decisiones tecnológicas tomadas y proporciona una estimación clara del coste de implementación del sistema propuesto.

Infraestructura Tecnológica

Descripción del entorno hardware y software empleado para la ejecución del proyecto. Toda infraestructura tecnológica requiere de unos recursos hardware y software para su ejecución. A continuación, detallaremos los equipos utilizados.

Software

Para el desarrollo de este trabajo todo el software que se ha utilizado es *opensource* o con licencias de libre uso, a excepción del sistema operativo del equipo de desarrollo que utiliza su licencia original de Windows y la de Office que es la proporcionada por la universidad.

Servidor

Para montar la infraestructura se ha contratado un servidor en el hosting *kimsufi* concretamente el modelo cuyo nombre comercial KS-B | Intel Xeon E5-1620v2 cuyas características son las siguientes:

Componente	**Especificación**
Procesador	Intel Xeon E5-1620v2
RAM	32 GB ECC 1333 MHz
Disco de sistema	1x120 GB SSD SATA
Sistema operativo	Ubuntu Server 24.04 “Noble Numbat”

Equipo de desarrollo

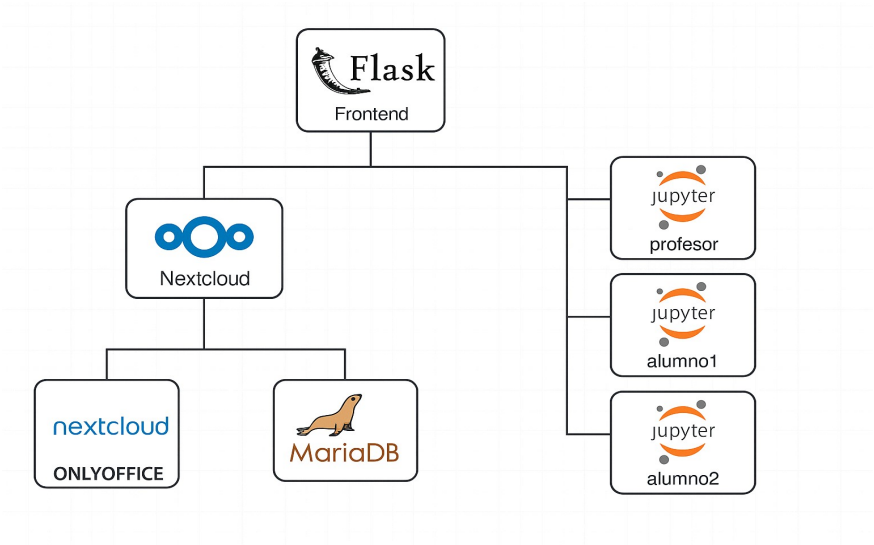
Para realizar el desarrollo se ha utilizado un equipo portátil personal para poder realizar la memoria y programar la plataforma, cuyas características son las siguientes:

Característica	**Descripción**
Modelo	**Acer Spin 5**
Procesador	**Intel Core i5 8250U (4 núcleos, 8 hilos, hasta 3.4 GHz)**
Memoria RAM	**8 GB DDR4**
Almacenamiento	**SSD de 256 GB (tipo SATA o NVMe, según configuración)**
Pantalla	**13.3″ táctil Full HD (1920 × 1080), panel IPS**

Gráficos	**Intel UHD Graphics 620 integrados**
Conectividad	**Wi-Fi 5 (802.11ac), Bluetooth 4.2**
Puertos	**2× USB 3.0, 1× USB-C, HDMI, jack de audio 3.5 mm**
Sistema operativo original	**Windows 10 Home**
Peso aproximado	**1,5 kg**
Características adicionales	**Bisagra 360° (modo tablet), pantalla táctil**

Arquitectura

Esquema de la arquitectura utilizada para crear la plataforma educativa:



Plataformas y Servicios Utilizados

Para la correcta utilización de los materiales didácticos preparados para este Trabajo Fin de Grado se ha desarrollado una plataforma software que pone a disponibilidad de los alumnos tanto los materiales formativos como sistemas de evaluación. De esta forma tienen a su disposición un sistema donde pueden trabajar sin depender de la configuración de sistemas y software de sus respectivos equipos.

La plataforma educativa se basa en el uso de microservicios soportados por Docker, para poder alojarlo es necesario un sistema operativo. Se decidió usar software no privativo, en este caso, un sistema operativo basado en Linux Debian. Concretamente el sistema operativo Ubuntu server 24.04 lts (long term support). Esta versión nos permite mantener la plataforma operativa durante un periodo de al menos 5 años con actualizaciones gratuitas, por lo que podríamos seguir hasta 2029 sin actualizar el sistema.

El servidor es accesible a través de la IP pública http://51.83.6.61

Para el desarrollo de la plataforma educativa, se ha adoptado un enfoque basado en microservicios, que permite una gestión más eficiente, escalable y modular de los distintos componentes tecnológicos implicados. Esta arquitectura facilita el mantenimiento del sistema y su evolución futura, permitiendo modificar o sustituir servicios de manera independiente.

Se ha utilizado Docker como sistema de contenerización para la implementación de los microservicios. Docker permite encapsular cada servicio (aplicación o componente) con todas sus dependencias, bibliotecas y configuración en contenedores aislados. Esto garantiza que cada parte del sistema se ejecute en un entorno controlado, predecible y replicable, sin interferencias entre los mismos.

Además, Docker facilita la orquestación de los diferentes servicios mediante redes virtuales internas, lo que simplifica la comunicación entre ellos y permite una implementación local o en la nube con mínimos ajustes.

Servicios desplegados en la plataforma educativa

Durante el desarrollo de la plataforma educativa he incorporado varios servicios clave y complementarios que permiten ampliar sus funcionalidades y mejorar la experiencia tanto del profesor como del alumno.

A continuación, se explican brevemente el propósito y la integración de cada uno de ellos.

Nextcloud

Uno de los pilares de la plataforma es Nextcloud. Utilizado como sistema de almacenamiento y compartición de archivos, esta herramienta ha permitido organizar las lecciones, ejercicios y documentos del curso de forma estructurada, así como gestionar los permisos de acceso para cada alumno o grupo. Gracias a sus funcionalidades adicionales, como los formularios integrados, y la posibilidad de trabajar con OnlyOffice, Nextcloud se convierte en un entorno muy completo para el intercambio de información educativa.

Para garantizar su correcto funcionamiento, ha sido desplegado como un contenedor independiente conectado con su base de datos (MariaDB) y con almacenamiento persistente.

JupyterLab

JupyterLab es el entorno de trabajo principal para los alumnos, una interfaz moderna para cuadernos Jupyter. Este servicio permite a los estudiantes desarrollar ejercicios interactivos de programación, análisis de datos y visualización, todo desde el navegador y sin necesidad de configurar nada en su equipo local. Su integración en la plataforma hace posible que los alumnos trabajen sobre los materiales que encuentran en Nextcloud y entreguen sus soluciones de forma centralizada y controlada.

Ha sido desplegado dentro de un contenedor Docker específico, con soporte para las principales bibliotecas de Python utilizadas en ciencia de datos (como pandas, matplotlib o scikit-learn).

Flask

Para facilitar el acceso inicial y centralizado a los distintos recursos, se creó un portal de entrada utilizando Flask, un *microframework* en Python. Este portal funciona como punto de acceso a los servicios desplegados y permite incluir elementos personalizados como enlaces, instrucciones de uso o avisos.

MariaDB

Uno de los requisitos técnicos de Nextcloud es disponer de una base de datos relacional. Para ello se optó por usar MariaDB, ya que es una robusta y compatible con MySQL que funciona bien en entornos contenerizados. La base de datos se despliega como un servicio aparte dentro del ecosistema de Docker y se encarga de gestionar la información interna de Nextcloud, como los usuarios, permisos, formularios o metadatos de los archivos.

OnlyOffice

Una de las funcionalidades más valoradas por los usuarios en entornos colaborativos es la posibilidad de visualizar y editar documentos directamente desde el navegador. Para lograr esto dentro de Nextcloud, integré el servicio de OnlyOffice, que permite abrir archivos de texto, hojas de cálculo y presentaciones sin necesidad de descargarlos. Esto resulta especialmente útil para los alumnos, ya que pueden consultar los materiales o rellenar formularios directamente desde la propia plataforma, y también para los profesores, que pueden editar las guías o subir correcciones sin abandonar el entorno de trabajo.

Costes Tecnológicos Estimados

A continuación, se presenta una estimación de los costes asociados al desarrollo y despliegue de la plataforma educativa. Esta tabla recoge tanto los recursos materiales como los servicios tecnológicos utilizados, teniendo en cuenta que el sistema se ha implementado mayoritariamente mediante software libre y contenedores Docker, lo que ha permitido minimizar los gastos asociados a licencias.

Los costes reflejados son aproximados y se han calculado considerando un entorno de pruebas funcional, destinado al uso académico, con posibilidad de escalar en función de la demanda o número de usuarios.

Concepto	Detalle	Importe (€)
Servidor VPS	9,90 €/mes × 12 meses + IVA	143.75 €
Equipo (Acer Spin 5)	Amortizado un año	140.00 €
Software utilizado	Software libre / código abierto	0.00 €
Ingeniero (desarrollo y soporte)	40 horas × 30 €/h	1200 €

Credenciales de acceso

A continuación, se detallan los puertos de acceso, credenciales y descripción de cada uno:

Servicio	Puerto externo	Usuario	Contraseña / Token
frontend	5000	—	—
db (MariaDB)	—	nextcloud	••••••••
nextcloud	8080	profesor	profesor
nextcloud	8080	alumno1	alumno1alumno1
nextcloud	8080	alumno2	alumno2alumno2
Nextcloud	8080	observadordocente	observadordocente
jupyteralumno1	8889	alumno1	alumno1
jupyteralumno2	8890	alumno2	alumno2
jupyterprofesor	8888	profesor	profesor
onlyoffice	9980	—	supersecret