AIEP
de la Universidad Andrés Bello

Taller de analítica en los negocios

Analítica en marketing

Escuela de Administración y Gestión Empresarial

Directora: Lorena Baus Piva

Elaboración

Experto disciplinar: Tomás von Bischoffshausen Gariazzo

Diseñador instruccional: Óscar González Cantin

Editor instruccional: David Villagrán Ruz

Validación

Experto disciplinar: Andrés Montecinos Rüth

Jefa de diseño instruccional: Sandra Betancur Cordero

Equipo de desarrollo

AIEP

AÑO

2025

Tabla de contenidos

Aprendizaje esperado de la semana

Aplican modelo de segmentación de consumidores con Kmeans a través de herramientas Python, considerando análisis de datos en el contexto de estrategias de marketing.

Fuente: Envato Elements (s. f.)

Introducción

¿Sabías que los datos que generamos diariamente pueden revelar patrones ocultos que transformen por completo las estrategias empresariales?

Te damos la bienvenida al módulo **Taller de analítica en los negocios**. En este módulo deberás poner en práctica técnicas avanzadas de análisis de datos para resolver problemas empresariales, mejorar campañas de marketing y optimizar decisiones estratégicas.

Este taller tiene la finalidad de que el(la) protagonista del aprendizaje seas tú, poniendo en práctica e integrando todos los conocimientos adquiridos durante la carrera, esta vez con foco en aplicaciones de negocio.

Si bien los apuntes servirán como recordatorios conceptuales de materias que ya has visto, serás tú quien tenga que resolver cómo implementar los códigos adecuados como un(a) verdadero(a) *data scientist*. Esto te permitirá tomar confianza en que puedes enfrentar proyectos nuevos por ti mismo(a) y buscar técnicas y códigos que te permitan resolver la tarea encomendada.

En esta primera semana, te adentrarás en los fundamentos del análisis de clúster y el modelamiento con algoritmos como KMeans. Exploraremos cómo limpiar y preparar datos para asegurar resultados confiables, cómo seleccionar el número óptimo de clústeres y cómo interpretar los grupos formados. Con un enfoque práctico, este apunte te guiará en el uso de herramientas de Python para implementar estas técnicas y traducir los resultados en estrategias accionables. Prepárate para comenzar tu viaje hacia la analítica avanzada.

1. Análisis del modelado de clúster para campañas de marketing

El análisis de clúster es una técnica en el ámbito del marketing para identificar patrones y segmentar consumidores en función de características específicas, de manera de construir grupos con características comunes e implementar estrategias personalizadas para cada uno, como preferencias de compra, hábitos de consumo, y rangos de ingresos, y así diseñar campañas más relevantes.

Uno de los enfoques recurrentes en las investigaciones es el uso del análisis de clúster para identificar grupos con comportamientos similares en campañas de marketing digital. Por ejemplo, algunos estudios han aplicado esta técnica para analizar la interacción de usuarios con contenidos en redes sociales, lo que les permite ajustar tanto el formato como el mensaje de sus publicaciones. Asimismo, la literatura destaca cómo los clústeres pueden revelar oportunidades de mercado al identificar nichos desatendidos.

2. Preparación de datos

Una etapa fundamental en el análisis de clúster es la **preparación de datos**, ya que la calidad y consistencia de los datos impactan directamente en la precisión y utilidad de los resultados del modelado. En este contexto, tres aspectos principales son esenciales: la limpieza de datos, la normalización/estandarización y el manejo de valores atípicos.

2.1. Limpieza de datos

La limpieza de datos es el proceso de **identificar y corregir errores o inconsistencias dentro del conjunto de datos**. Este paso asegura que los datos utilizados para el análisis sean precisos y confiables mediante las siguientes acciones:

- Identificar y remover registros repetidos para evitar que los clústeres se vean influenciados por datos redundantes.
- Imputación o eliminación de registros incompletos si el impacto es marginal.
- Corregir formatos inconsistentes (como fechas y cadenas de texto) y garantizar que las variables tengan el tipo de dato correcto para el análisis.

2.2. Normalización/estandarización de datos

Normaliza o estandariza las variables para evitar que las escalas diferentes sesguen el análisis. Esto asegura que todas las características tengan igual peso en la formación de clústeres.

Usa normalización cuando los valores deben estar en un rango específico. y estandarización si los algoritmos que vas a utilizar son sensibles a distribuciones normales (como KMeans).

2.3. Manejo de valores atípicos

Identifica y trata los valores que se desvían significativamente del resto, para evitar que afecten negativamente el modelo. Usa un análisis gráfico como *boxplots* o un análisis de z-scores para detectar valores atípicos

numéricamente y elimina valores extremos para que el modelo no se distorsione.

Ser tratados con técnicas específicas, como el recorte de valores extremos (winsorización).

El manejo adecuado de valores atípicos mejora la precisión del análisis y asegura que los clústeres reflejen patrones reales y no anomalías en los datos.

3. Recordando KMeans

Como viste en detalle en cursos pasados, el **modelamiento de clustering con KMeans** es uno de los enfoques más comunes y efectivos para realizar análisis de segmentación. En esta sección, revisaremos conceptos clave y los pasos prácticos para implementar el algoritmo KMeans, desde la inicialización de centroides hasta la determinación del número óptimo de clústeres utilizando el método del codo. Recordemos:

- KMeans es un algoritmo de agrupamiento no supervisado que divide los datos en *k* clústeres, asignando cada punto de datos al clúster más cercano basado en la distancia a los centroides. Este método minimiza la suma de las distancias cuadradas entre los puntos y el centroide de su clúster.
- La inicialización de centroides es un paso crítico, ya que los centroides iniciales afectan el desempeño y los resultados del algoritmo KMeans. Una inicialización pobre puede llevar a soluciones subóptimas.

- El algoritmo KMeans termina cuando los centroides dejan de moverse significativamente entre iteraciones, o cuando se alcanza un número máximo de iteraciones.
- Elegir el número adecuado de clústeres es fundamental para garantizar que los grupos sean significativos y útiles para el análisis.
- El método del codo evalúa la inercia (suma de las distancias cuadradas dentro de los clústeres) para diferentes valores de *k*. A medida que aumenta el número de clústeres, la inercia disminuye, pero la mejora marginal disminuye después de cierto punto, formando un 'codo' en el gráfico.

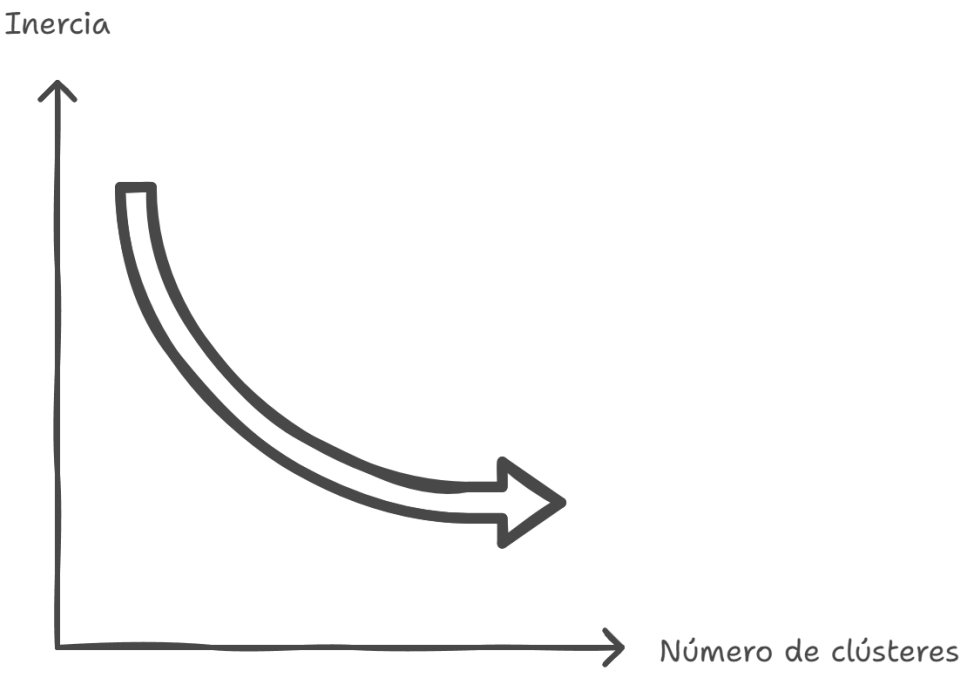


Figura 1. Relación entre el número de clústeres y la inercia en el método del codo. La figura ilustra cómo la inercia disminuye a medida que aumenta el número de clústeres, permitiendo identificar el punto óptimo donde se alcanza un equilibrio entre complejidad y representatividad de los clústeres

4. Evaluación del modelo

Una vez que se ha realizado el modelado de clúster, es fundamental evaluar la calidad del agrupamiento para garantizar que los clústeres formados sean significativos y representen bien los datos. Recordemos:

4.1. Coeficiente de silhouette

El coeficiente de silhouette mide qué tan bien se agrupan los puntos dentro de un clúster y qué tan separados están de otros clústeres. Este valor oscila entre -1 y 1:

- Valores cercanos a 1 indican que los puntos están bien agrupados.
- Valores cercanos a 0 sugieren que los puntos están en el límite entre clústeres.
- Valores negativos indican que los puntos probablemente estén mal asignados.

4.2. Criterios de evaluación interna: SSE e inercia

La métrica SSE (*Sum of Squared Errors*) es la suma de los errores cuadráticos dentro de cada clúster, representando qué tan cerca están los puntos de sus centroides. Un SSE bajo indica clústeres compactos.

Por otro lado, la métrica SSE mide la variación dentro de los clústeres, calculando la suma de las distancias al cuadrado entre los puntos y sus centroides.

Ambas métricas (coeficiente de silhouette y SSE/inercia) ofrecen información complementaria:

- El coeficiente de silhouette evalúa la calidad del agrupamiento, considerando la separación entre clústeres.
- La inercia (o SSE) se utiliza para determinar la compacidad de los clústeres y es especialmente útil en la elección del número óptimo de clústeres mediante el método del codo.

4.3. Aplicación práctica combinada

- **Determina el número óptimo de clústeres**: usa el método del codo con la inercia para encontrar un rango razonable de clústeres posibles.
- **Evalúa la calidad del agrupamiento**: calcula el coeficiente de silhouette para los números de clústeres seleccionados.
- **Selecciona el mejor modelo**: escoge la configuración con un equilibrio entre un coeficiente de silhouette alto y una inercia baja.

5. Clústeres

Una vez creados los clústeres, es fundamental **interpretarlos** para comprender sus características principales, **visualizarlos** para confirmar su coherencia y **analizar los factores dominantes que definen cada grupo**. Esta etapa transforma el análisis matemático en información práctica.

5.1. Interpretación de clústeres

La interpretación de clústeres implica **identificar patrones dentro de los grupos formados**, como características comunes entre los puntos de datos de cada clúster. Este análisis responde a preguntas como:

- ¿Qué define a cada clúster? (por ejemplo, características demográficas, preferencias o comportamientos).
- ¿En qué se diferencian los clústeres entre sí?

Esto muestra el promedio de cada característica en cada clúster, ayudándote a identificar qué variables destacan.

Las características dominantes son aquellas que tienen valores consistentes y significativos dentro de un clúster, distinguiéndose de otros grupos.

Para identificarlas:

- Calcula las medias y medianas de cada característica por clúster (como en la sección de interpretación).
- Identifica variables con diferencias marcadas entre los clústeres.
- Usa la importancia de características (si el modelo soporta esta función, como Random Forest o KMeans avanzado).

Una descripción de las variables más relevantes que definen cada clúster, permite generar *insights* prácticos como segmentación de clientes o identificación de nichos de mercado.

6. Perfilado de consumidores

El perfilado de consumidores por clúster es una etapa esencial para transformar el análisis de datos en información práctica y aplicable. Consiste en describir las características comunes de los consumidores en cada clúster, como edad, ingresos, preferencias o cualquier otra variable relevante. Este paso permite a las organizaciones desarrollar estrategias

personalizadas basadas en la composición de cada grupo. El perfilado busca responder preguntas clave sobre cada clúster, como:

- ¿Qué características demográficas definen al clúster? (por ejemplo, edad promedio o nivel de ingresos).
- ¿Qué comportamientos o preferencias destacan? (como preferencias de productos o frecuencia de compra).
- ¿Cómo se diferencian estos consumidores de los de otros clústeres?
- Este análisis ayuda a entender qué hace único a cada grupo y cómo puede ser segmentado para actividades específicas, como campañas de marketing o diseño de productos.

Para lograr lo anterior:

- Agrupa los datos según los clústeres formados. Una vez que los datos estén etiquetados con los clústeres generados, puedes agruparlos para calcular estadísticas descriptivas.
- Busca variables con diferencias significativas entre clústeres (por ejemplo, un clúster con ingresos notablemente altos).
- Observa qué valores son consistentes dentro de un clúster.

Por ejemplo:

Supongamos que has realizado un análisis de clústeres con datos de consumidores que incluyen su edad, ingresos y preferencias de productos. Tras agrupar los datos, obtienes las siguientes estadísticas descriptivas:

Clúster	Edad promedio	Ingresos promedio (USD)	Preferencia de producto
0	25	20,000	Tecnología
1	40	50,000	Productos de lujo
2	60	30,000	Salud y bienestar

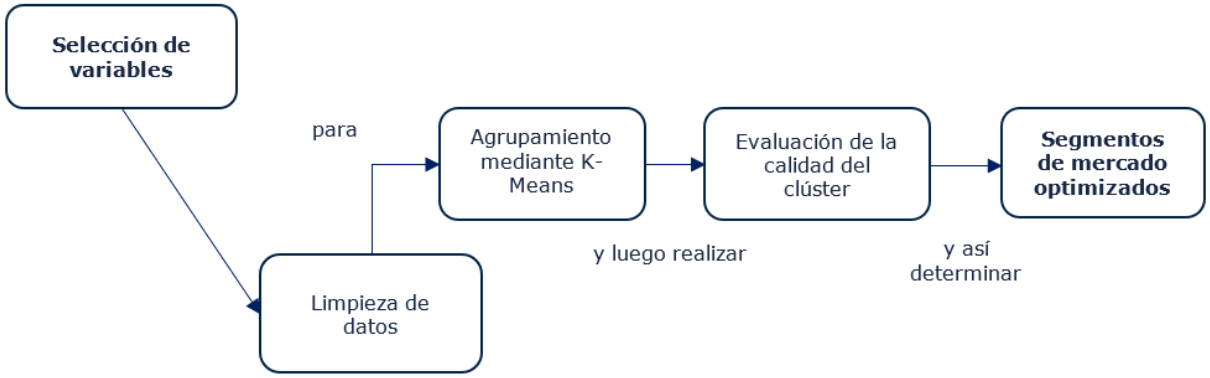
Interpretación:

- **Clúster 0:** consumidores jóvenes con ingresos bajos, interesados principalmente en tecnología.
- **Clúster 1:** adultos de mediana edad con ingresos altos, atraídos por productos de lujo.
- **Clúster 2:** personas mayores con ingresos moderados, centradas en salud y bienestar.

Este perfilado puede orientar decisiones como crear campañas específicas para cada grupo u ofrecer promociones personalizadas según el segmento.

Cierre

Por medio del siguiente organizador gráfico se destacan las ideas clave de esta semana:



La preparación de datos y el modelamiento de clústeres son pasos fundamentales en el análisis de segmentación, que permiten identificar patrones significativos en grandes volúmenes de datos. Mediante técnicas como la limpieza, la normalización y el uso del algoritmo KMeans, puedes agrupar consumidores de manera precisa, generando información clave para decisiones estratégicas. Además, la interpretación de los clústeres, respaldada por métricas de evaluación y visualizaciones claras, transforma los datos en conocimiento accionable.

Referencias

Envato Elements (s. f.). Visión desde un ángulo elevado de los trabajadores. [Imagen]. https://elements.envato.com/es/high-angle-view-on-working-people-VK423DJ

(s. f.) Un hombre sostiene una caja de cartón y está de pie frente a una pizarra blanca. Está haciendo una presentación y sostiene la caja para mostrar algo [Imagen]. https://elements.envato.com/es/group-of-young-male-and-female-business-people-in--EUNHC8M

Las figuras (diagramas e imágenes) utilizadas en este apunte fueron generadas mediante Napkin (https://www.napkin.ai/)