2. APRENDIZAJE NO SUPERVISADO

Imagina que tienes una gran caja llena de piezas de rompecabezas, pero no tienes la imagen de referencia. Tu tarea es organizar las piezas en grupos, basándote únicamente en sus formas y colores. Así es como funciona el aprendizaje no supervisado: se trata de encontrar patrones y estructuras en datos sin tener una "respuesta correcta" predeterminada.

En este tipo de aprendizaje, el algoritmo recibe un conjunto de datos sin etiquetas o categorías asignadas. El objetivo es descubrir agrupaciones, relaciones o características interesantes que no eran evidentes a simple vista. No se le dice al algoritmo qué buscar exactamente, solo se le permite explorar los datos.

Desde el punto de vista matemático, el aprendizaje no supervisado trabaja sobre espacios multidimensionales, donde cada dato puede ser representado como un punto en un gráfico. Estos puntos pueden tener múltiples dimensiones, dependiendo del número de características que describen cada observación. Por ejemplo, en el caso de nuestras piezas de rompecabezas, cada pieza podría representarse por características como forma, color y tamaño.

Los algoritmos de aprendizaje no supervisado utilizan conceptos estadísticos para identificar patrones. Uno de estos conceptos es la medida de distancia, que ayuda a determinar qué tan cerca o lejos están dos datos en el espacio. Entre las medidas más comunes se encuentra la distancia euclidiana, que en términos sencillos calcula la distancia "en línea recta" entre dos puntos.

Otro concepto fundamental es el de la varianza, que mide cuán dispersos están los datos alrededor de un punto central, como la media. En el aprendizaje no supervisado, entender la varianza puede ayudar a identificar qué tan bien se agrupan los datos y si existen subgrupos diferenciados dentro de un conjunto de datos.

El aprendizaje no supervisado también puede involucrar técnicas de reducción de dimensionalidad. Estas técnicas, como el Análisis de Componentes Principales (PCA), buscan simplificar la representación de los datos, reduciendo el número de variables a aquellas que capturan la mayor parte de la información. Esto permite visualizar mejor la estructura subyacente de los datos en un espacio más reducido.

El proceso de identificar patrones sin una guía externa se basa en encontrar similitudes entre los datos. Por ejemplo, si observamos las piezas del rompecabezas, podemos agrupar aquellas que tienen colores y formas similares. Matemáticamente, esto se traduce en buscar grupos de puntos que se encuentren más próximos entre sí en el espacio multidimensional.

Una vez que se identifican estos grupos o patrones, se pueden extraer conclusiones sobre la estructura interna del conjunto de datos. Este análisis permite entender mejor la distribución de los datos y puede ser la base para tomar decisiones informadas, como diseñar estrategias o detectar anomalías.

El aprendizaje no supervisado es especialmente útil cuando no se dispone de información previa o etiquetas sobre los datos. Al permitir que los datos "hablen por sí mismos", se descubren relaciones y agrupaciones que de otra forma pasarían desapercibidas, brindando una nueva perspectiva sobre el problema a resolver.

En resumen, el aprendizaje no supervisado es una herramienta poderosa en ciencia de datos que, mediante el uso de conceptos matemáticos y estadísticos, permite explorar y entender conjuntos de datos complejos. Sin necesidad de etiquetas o respuestas predeterminadas, esta técnica revela estructuras y patrones ocultos, ayudando a transformar datos brutos en conocimiento útil.

Al no disponer de observaciones etiquetadas, el proceso o **fases** del aprendizaje no supervisado es diferente al del caso supervisado. Vamos a verlo.

Imagina que volvemos a nuestra caja de rompecabezas sin imagen de referencia. En este caso, el proceso de aprendizaje no supervisado se puede dividir en varias fases, cada una de las cuales contribuye a descubrir la estructura oculta de las piezas. Estas fases son la configuración, el entrenamiento, la prueba y la inferencia.

• En la **fase de configuración**, establecemos el escenario para nuestro experimento. Es como preparar la mesa de trabajo: definimos las características que observaremos en cada pieza, elegimos el algoritmo que nos ayudará a agrupar las piezas y ajustamos parámetros iniciales, como la forma de medir la similitud entre ellas. Este paso es crucial, ya que una buena configuración sentará las bases para un análisis exitoso.

Durante la configuración, también se decide qué métricas se utilizarán para evaluar el desempeño del agrupamiento. Aunque no se tengan etiquetas que confirmen la agrupación, se pueden definir criterios internos, como la cohesión dentro de cada grupo o la separación entre grupos, que nos indicarán si la estructura encontrada tiene sentido.

• La **fase de entrenamiento** es donde el algoritmo se pone a trabajar. En nuestro ejemplo, es el momento en que se empieza a examinar cada pieza y a agruparlas en función de sus características. Aquí el algoritmo analiza la distribución de las piezas en el espacio definido por sus atributos, buscando patrones y similitudes sin tener una respuesta correcta preestablecida.

Mientras el algoritmo "entrena", va ajustando sus parámetros internamente para que las agrupaciones sean lo más homogéneas posible. Este proceso puede implicar iteraciones en las que se reagrupan las piezas, se refinan los límites entre grupos y se optimiza la forma en que se mide la cercanía entre ellas.

• Una vez que el algoritmo ha finalizado su entrenamiento, llega la **fase de prueba**. En este paso se evalúa la solidez de las agrupaciones encontradas. Se analizan medidas como la consistencia interna de cada grupo o la separación efectiva entre grupos. Es un poco como revisar la caja de rompecabezas y asegurarse de que las piezas agrupadas realmente tienen características similares.

La prueba nos permite identificar si el algoritmo ha sido capaz de descubrir patrones significativos o si es necesario ajustar la configuración o incluso el algoritmo mismo. Aunque no se disponga de etiquetas para comparar, estas evaluaciones internas son vitales para validar el trabajo realizado durante el entrenamiento.

• La **fase de inferencia** es el momento en que ponemos en práctica lo aprendido. Con las reglas y patrones extraídos del entrenamiento, el modelo ahora puede asignar nuevas piezas a los grupos ya identificados. Es similar a recibir nuevas piezas de rompecabezas y, basándonos en lo que aprendimos, decidir a cuál grupo pertenece cada una.

Durante la inferencia, el modelo actúa como una guía que, sin intervención humana directa, clasifica o agrupa nuevas observaciones según las similitudes aprendidas previamente. Este paso demuestra la capacidad del modelo para generalizar y aplicar su conocimiento a nuevos datos, lo cual es esencial en cualquier aplicación práctica.

En conjunto, estas fases –configuración, entrenamiento, prueba e inferencia– forman un ciclo en el aprendizaje no supervisado. Cada etapa es fundamental para descubrir y validar patrones en conjuntos de datos sin etiquetas, permitiendo que los datos "hablen por sí mismos" y revelen su estructura interna.

Así, a través de este proceso, el aprendizaje no supervisado se convierte en una herramienta poderosa para explorar y entender grandes volúmenes de datos, organizándolos de manera que se descubran relaciones y agrupamientos que, de otra forma, pasarían desapercibidos.

2.1. SEGMENTACIÓN

Imagina que tienes un gran mercado con miles de personas y tu objetivo es entender quiénes son, qué les interesa y cómo puedes atender mejor sus necesidades. Los métodos de segmentación actúan como una lupa que te permite dividir ese mercado en grupos homogéneos, de modo que puedas crear estrategias específicas para cada uno. Es como organizar una gran fiesta y, en lugar de tratar a todos de la misma manera, decides agrupar a los invitados según sus gustos, edades o intereses, para que cada grupo tenga una experiencia personalizada.

En el mundo de los negocios, la segmentación es fundamental para adaptar productos y servicios a las preferencias de diferentes públicos. Por ejemplo, una empresa de moda puede utilizar estos métodos para distinguir entre clientes que buscan ropa casual y aquellos que prefieren atuendos formales, permitiéndole ofrecer colecciones específicas que respondan a cada estilo.

Los métodos de segmentación se basan en diversas técnicas analíticas y estadísticas. Algunas utilizan algoritmos de agrupamiento, como el K-means o el clustering jerárquico, que identifican de forma automática grupos de clientes basándose en patrones en los datos. Estos algoritmos actúan como un sistema inteligente que detecta similitudes entre las personas, sin necesidad de que previamente se haya definido a qué grupo pertenece cada individuo.

Otros métodos de segmentación pueden apoyarse en análisis de factores o en técnicas de reducción de dimensionalidad, que simplifican la información y resaltan las variables más relevantes. Esto es similar a filtrar una imagen borrosa para ver claramente los rasgos más distintivos de cada grupo de personas.

La utilidad de estos métodos no se limita solo al ámbito comercial. En el **sector de la salud**, por ejemplo, se utilizan para identificar grupos de pacientes con características similares, lo que puede ayudar a personalizar tratamientos o identificar patrones de enfermedad en poblaciones específicas. De esta forma, se pueden optimizar recursos y mejorar la calidad de la atención médica.

En el **sector financiero**, la segmentación es clave para evaluar el riesgo y ofrecer productos financieros adecuados. Al agrupar a los clientes según su comportamiento crediticio, ingresos o historial de inversiones, las instituciones pueden diseñar estrategias de marketing más eficientes y mitigar riesgos, asignando recursos de manera más inteligente.

El **ámbito del turismo** también se beneficia de estos métodos. Las agencias de viajes pueden segmentar a los clientes por preferencias de destino, tipo de viaje o presupuesto, lo que les permite crear paquetes turísticos a la medida y mejorar la satisfacción de los viajeros.

Incluso en la **educación**, la segmentación juega un papel importante. Las instituciones educativas pueden utilizarla para identificar estilos de aprendizaje o necesidades específicas de los estudiantes, facilitando la creación de programas de enseñanza personalizados y mejorando los resultados académicos.

En resumen, los métodos de segmentación son herramientas poderosas que ayudan a dividir grandes volúmenes de datos en grupos significativos y homogéneos. Esto permite no solo entender mejor a los clientes o usuarios, sino también diseñar estrategias y soluciones adaptadas a las necesidades específicas de cada grupo, haciendo que la interacción y el servicio sean mucho más efectivos.

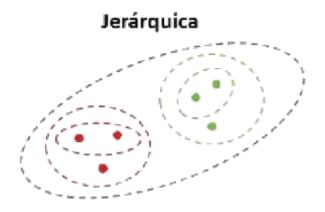
Estos métodos se aplican en múltiples sectores, desde el comercio y la salud hasta las finanzas, el turismo y la educación, demostrando su versatilidad y relevancia en cualquier campo donde comprender a la audiencia sea clave para el éxito.

Tipos de Segmentaciones

Imagina que estás organizando una gran reunión en la que deseas agrupar a los invitados según sus intereses. Los distintos métodos de segmentación son como estrategias para dividir a estos invitados en grupos, y cada una tiene su propio "estilo" para hacerlo. Te explicaré cada uno, acompañado de un dibujo sencillo para que lo visualices.

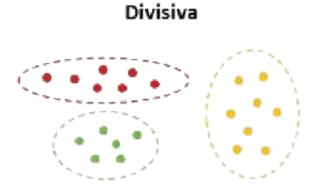
Segmentación Jerárquica:

Piensa en un árbol genealógico en el que, a partir de un grupo inicial, se crean ramas que se dividen en subgrupos cada vez más específicos. Es un proceso ordenado en el que cada nivel revela una segmentación más fina de los datos.



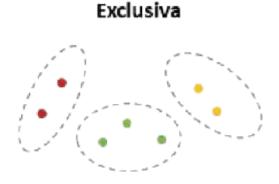
Segmentación Divisiva:

Este método es como partir una gran tarta: empiezas con un grupo único y lo vas dividiendo en partes, de manera secuencial y descendente. Cada división intenta separar lo más diferente posible los subgrupos, generando una estructura similar a la jerárquica, pero siempre desde el enfoque de dividir el grupo completo en partes más pequeñas.



Segmentación Exclusiva:

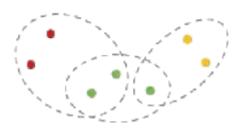
Imagina que cada invitado debe elegir una única mesa para sentarse y no puede ir a más de una. En la segmentación exclusiva, cada individuo se asigna a un solo grupo sin superposición. Esto garantiza que los grupos sean claros y sin ambigüedades.



Segmentación Superpuesta:

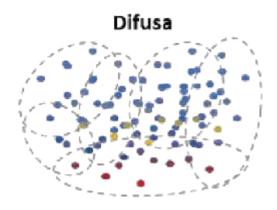
Ahora imagina que algunos invitados tienen intereses variados y pueden pertenecer a más de una mesa. La segmentación superpuesta permite que un mismo individuo se incluya en varios grupos, reflejando la realidad de que las personas pueden tener múltiples características o intereses.

Superpuesta



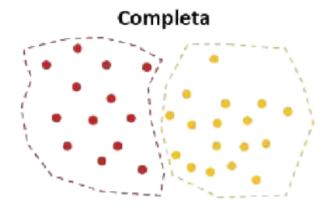
Segmentación Difusa (Fuzzy):

En este caso, en lugar de asignar a cada invitado a una única mesa, se les da una "probabilidad" o grado de pertenencia a cada grupo. Por ejemplo, un invitado puede tener un 70% de pertenencia a un grupo y un 30% a otro, reflejando la ambigüedad y solapamiento en sus intereses.



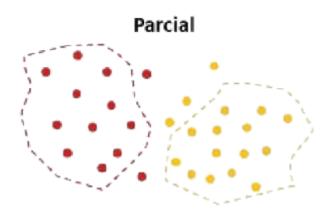
Segmentación Completa:

Este método asegura que cada invitado es asignado a algún grupo, de manera que no queda nadie sin clasificar. Es un enfoque en el que se cubre el total de la población, generando grupos que, en conjunto, abarcan a todos los individuos.



Segmentación Parcial:

Por último, en la segmentación parcial, no todos los individuos tienen que pertenecer a un grupo. Algunos pueden quedar sin asignar o solo algunos subgrupos se definen para ciertos segmentos de la población. Esto es útil cuando solo una parte de los datos muestra patrones claros, dejando otros sin clasificar.



Cada uno de estos métodos se utiliza según las necesidades del análisis y las características de los datos. La elección dependerá de si buscas una asignación clara y exclusiva, una representación más realista con solapamientos, o incluso una estructura jerárquica que permita ver relaciones en distintos niveles. En sectores como el marketing, la salud o las finanzas, estos enfoques ayudan a descubrir patrones y a personalizar estrategias, haciendo que la segmentación sea una herramienta fundamental para la toma de decisiones.

Estos relatos y dibujos sencillos buscan ilustrar cómo cada técnica se adapta a distintos escenarios, ayudando a visualizar la esencia de cada método de segmentación. Así, desde una simple división de invitados en mesas hasta la asignación de probabilidades, cada técnica permite comprender mejor la diversidad y complejidad de los datos.

Clasificación de las métricas de Segmentación.

A continuación, se muestra una tabla que agrupa en cuatro familias algunos de los algoritmos más representativos de las técnicas de segmentación:

Familia	Algoritmos Típicos
Basadas en la conectividad Métricas de distancia	Algoritmo de crecimiento de regiones, Watershed, Segmentación basada en grafos, Componentes conexas.
Basadas en el prototipo Métricas de distancia	K-means, K-medoids, Self-Organizing Maps (SOM), Learning Vector Quantization (LVQ).
Basadas en la distribución Métricas de probabilidad	Modelos de mezcla gaussiana (GMM), Algoritmo de Expectation- Maximization (EM), Clustering bayesiano.
Basadas en la concentración Métricas de densidad	DBSCAN, OPTICS, Mean-Shift, HDBSCAN.

Esta clasificación permite entender cómo cada familia aborda la segmentación de datos según criterios como la conectividad entre elementos, la representación a través de prototipos, la asunción de un modelo estadístico subyacente en la distribución o la densidad de puntos en el espacio.

2.2. REGLAS DE ASOCIACIÓN Y PATRONES SECUENCIALES

Reglas de Asociación

Imagina que entras en una tienda y observas cómo se comportan los clientes mientras compran. Sin que nadie les diga qué productos comprar, algunos artículos parecen ir juntos de manera recurrente. Las reglas de asociación en el aprendizaje no supervisado son justamente esa herramienta que descubre patrones ocultos y relaciones entre los elementos, sin una guía previa.

En este escenario, las **reglas de asociación** buscan identificar conjuntos de productos que tienden a comprarse conjuntamente. Por ejemplo, al analizar las compras, el algoritmo puede encontrar que cuando alguien adquiere pan, es muy probable que también lleve mantequilla. Este hallazgo se expresa en una regla del tipo "pan → mantequilla", que indica una relación de coocurrencia entre estos productos.

Desde el punto de vista matemático, estas reglas se fundamentan en conceptos como el soporte, la confianza y el lift. El soporte mide la frecuencia con la que aparece una combinación de productos en todas las transacciones, y es fundamental para determinar la relevancia de la asociación en el conjunto de datos. Imagina que, de cada 100 transacciones, 20 incluyen tanto pan como mantequilla; ese 20% representa el soporte de la regla.

La **confianza** es otro concepto clave. Esta métrica indica la probabilidad de que se compre mantequilla cuando se ha comprado pan. Si de los 30 clientes que compran pan, 15 también compran mantequilla, la confianza de la regla "pan → mantequilla" sería del 50%. Es una forma de medir la fiabilidad de la asociación entre los productos.

El **lift**, por su parte, evalúa la fortaleza de la relación comparando la confianza de la regla con la probabilidad de que ambos productos se compren de manera independiente. Si el lift es mayor que 1, significa que la ocurrencia conjunta de pan y mantequilla es más frecuente de lo que se esperaría por azar, lo que resalta una relación significativa entre ellos.

Estas reglas no solo se utilizan en tiendas; su utilidad se extiende a diversos ámbitos. Por ejemplo, en el ámbito digital, un sitio web de streaming podría usar reglas de asociación para recomendar películas, al identificar que usuarios que ven una película de acción tienden a disfrutar también de una comedia de situación. Así, el sistema sugiere contenido relacionado sin intervención humana directa.

Otro ejemplo puede ser en la optimización de inventarios. Las reglas de asociación ayudan a prever qué productos deberían estar ubicados cerca en una tienda para facilitar la experiencia del cliente. De este modo, si se detecta que los clientes que compran café también adquieren azúcar, estos productos se pueden ubicar juntos para mejorar el flujo de compra.

El proceso de generar estas reglas implica, en primer lugar, la recolección y preprocesamiento de los datos. Se analizan las transacciones registradas, y el algoritmo, a través de técnicas como el algoritmo Apriori, identifica todas las combinaciones de productos que superan ciertos umbrales de soporte y confianza. Este método sistemático permite extraer las reglas más relevantes del vasto conjunto de datos.

Una vez extraídas, estas reglas deben ser evaluadas y refinadas para asegurarse de que sean útiles y accionables. Las empresas pueden ajustar los umbrales de soporte y confianza según sus necesidades y la dinámica de sus mercados, asegurando así que las recomendaciones o acciones derivadas de las reglas sean realmente valiosas.

En resumen, las reglas de asociación son una potente herramienta del aprendizaje no supervisado, que, sin necesidad de etiquetas, revela patrones interesantes en los datos. Permiten descubrir relaciones inesperadas y tomar decisiones informadas, desde mejorar estrategias de venta hasta optimizar la experiencia del cliente.

Considera un conjunto de ítems I y un conjunto de transacciones T, donde cada transacción $t \in T$ es un subconjunto de I. Una regla de asociación se expresa en la forma:

$$X \to Y, X \cap Y = \emptyset$$

donde X y Y son conjuntos disjuntos de items denominados itemsts. X recibe el nombre de **antencedente** (o cuerpo) de la regla, mientras que Y es el **consecuente** (o cabeza)

Soporte

El **soporte** de la regla $X \to Y$ es la proporción de transacciones que contienen ambos conjuntos X e Y. Se define como:

$$\operatorname{soporte}(X o Y) = \frac{|\{t \in T \mid X \cup Y \subseteq t\}|}{|T|}$$

Este valor indica cuán frecuente es la aparición conjunta de X e Y en el conjunto de transacciones.

Confianza

La **confianza** de la regla mide la probabilidad de que Y aparezca en una transacción dado que X ya está presente. Se define como:

$$\operatorname{confianza}(X \to Y) = \frac{\operatorname{soporte}(X \to Y)}{\operatorname{soporte}(X)} = \frac{|\{t \in T \mid X \cup Y \subseteq t\}|}{|\{t \in T \mid X \subseteq t\}|}$$

Esto indica qué tan confiable es la regla, es decir, si X ocurre, ¿con qué frecuencia ocurre también Y?

Lift

El **lift** evalúa la fuerza de la asociación entre X e Y comparando la confianza de la regla con la probabilidad de que Y ocurra de forma independiente. Se define como:

$$\operatorname{lift}(X \to Y) = \frac{\operatorname{confianza}(X \to Y)}{\operatorname{soporte}(Y)}$$

Un lift mayor que 1 indica que la presencia de X incrementa la probabilidad de que Y ocurra, evidenciando una relación positiva entre los dos conjuntos.

En conjunto, estos tres parámetros—soporte, confianza y lift—forman el formalismo matemático básico que determina la validez y la relevancia de una regla de asociación dentro del aprendizaje no supervisado.

Patrones Secuenciales

Imagina que en lugar de analizar simplemente qué productos se compran juntos, te interesa saber en qué orden se realizan las compras. Por ejemplo, en una tienda online, podrías descubrir que muchos clientes primero visitan la página de ofertas, luego exploran la sección de electrónica y, finalmente, adquieren algún accesorio para móviles. Estos recorridos de comportamiento se conocen como **patrones secuenciales**.

En el contexto del aprendizaje no supervisado, los patrones secuenciales identifican secuencias o cadenas de eventos que ocurren en un orden determinado a lo largo del tiempo. Cada secuencia se compone de elementos o ítems que se suceden, y el objetivo es descubrir relaciones temporales que no son evidentes al analizar los datos de forma aislada.

Matemáticamente, podemos representar una secuencia como una lista ordenada de conjuntos, por ejemplo:

$$\langle X_1, X_2, X_3, \dots, X_n \rangle$$

donde cada X_i es un conjunto de ítems que ocurren en un instante o evento específico. Un patrón secuencial es una secuencia que aparece en los datos con una frecuencia suficiente, superando un umbral mínimo de soporte.

Para extraer estos patrones, se analizan los datos de transacciones ordenadas temporalmente. Por ejemplo, si tenemos la secuencia de compras de varios clientes, un algoritmo puede identificar que la secuencia $\langle \{A\}, \{B\} \rangle$ ocurre frecuentemente, lo que indicaría que después de comprar el producto A, es común que el cliente adquiera el producto B.

Este tipo de análisis es especialmente útil en ámbitos como el marketing, donde entender la evolución del comportamiento del cliente a lo largo del tiempo permite diseñar estrategias personalizadas. Por ejemplo, un minorista podría utilizar estos patrones para recomendar productos basándose en la secuencia de navegación o compra del usuario.

Además, los patrones secuenciales no solo ayudan a predecir qué productos se podrían comprar en el futuro, sino que también permiten optimizar procesos y mejorar la experiencia del usuario. Conocer la secuencia habitual de acciones puede guiar la organización del contenido en un sitio web o la disposición de productos en una tienda física.

En resumen, los patrones secuenciales en el aprendizaje no supervisado son la extensión de las reglas de asociación, pero que además capturan el factor tiempo y orden en el que ocurren los eventos. Permiten descubrir cómo se encadenan las acciones y cómo un evento puede influir en los que le siguen, ofreciendo una perspectiva más dinámica y detallada del comportamiento dentro de un sistema.

Estos análisis, basados en métodos matemáticos y algoritmos especializados, ayudan a transformar datos ordenados en conocimiento práctico, facilitando la toma de decisiones en diversos sectores, desde el comercio digital hasta la optimización de procesos industriales.

La siguiente tabla reúne los principales algoritmos y aplicaciones para el análisis de asociaciones.

A continuación, te presento una tabla que agrupa algunos de los algoritmos más representativos para la generación de reglas de asociación y la extracción de patrones secuenciales, junto con ejemplos para facilitar su comprensión:

Tipo de Técnica	Algoritmo	Descripción Breve	Ejemplo
Reglas de Asociación	Apriori	Explora el espacio de ítems de manera iterativa, generando candidatos y eliminando aquellos que no cumplen el soporte mínimo.	Descubrir que en un supermercado, el 20% de las transacciones donde se compra pan también incluyen mantequilla.
	Eclat	Utiliza un enfoque basado en la intersección de listas de transacciones para extraer conjuntos frecuentes de ítems.	Identificar grupos de productos (por ejemplo, café, azúcar y leche) que aparecen juntos con frecuencia en las compras de los clientes.

Tipo de Técnica	Algoritmo	Descripción Breve	Ejemplo
	FP-Growth	Construye un árbol compacto (FP-tree) para evitar generar candidatos en exceso y extraer las reglas de manera más eficiente.	Extraer patrones de compra en grandes bases de datos, como que los clientes que compran frutas frescas también tienden a comprar verduras.
Patrones Secuenciales	GSP (Generalized Sequential Pattern)	Encuentra secuencias frecuentes en datos ordenados temporalmente, considerando restricciones en la longitud y tiempo entre ítems.	Descubrir que muchos clientes visitan primero la sección de ofertas y luego la de electrónica antes de comprar accesorios.
	SPADE	Emplea técnicas de partición y equivalencia para descubrir patrones secuenciales, aprovechando la estructura de los datos.	Identificar que en una plataforma de streaming, los usuarios que ven una serie de acción suelen ver una comedia en la siguiente sesión.
	PrefixSpan	Se centra en proyectar la base de datos en prefijos frecuentes, extrayendo secuencias sin necesidad de generar candidatos globales.	Extraer patrones en la navegación web, como que los usuarios que comienzan en la página principal tienden a visitar la sección de productos y luego el carrito.

Esta tabla resume los principales algoritmos asociados a cada técnica, resaltando tanto su enfoque metodológico como ejemplos prácticos de su aplicación en contextos reales.