Las tareas del machine learning no supervisado

En la Fig. 1 se pueden apreciar las tareas típicas del aprendizaje no supervisado. Veamos en más detalle cada una de ellas:

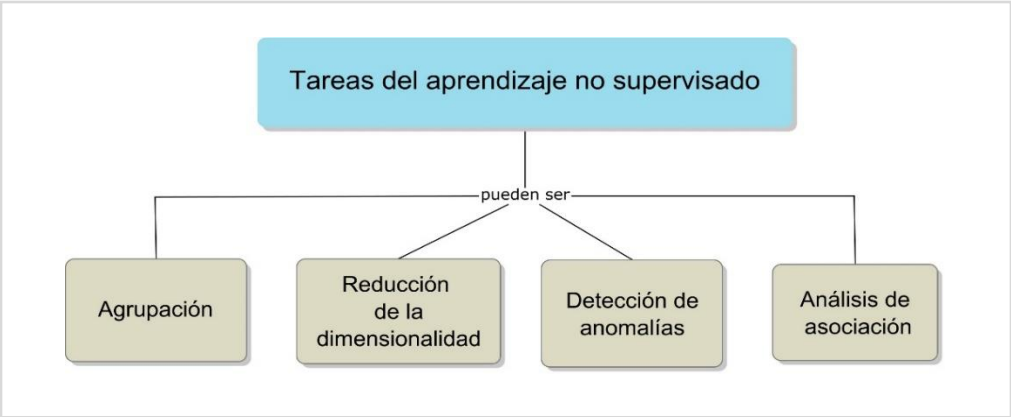

Fig. 1. Tareas del aprendizaje no supervisado.

- **Agrupación**.

La agrupación, también conocida como segmentación de datos, es, quizás, la tarea descriptiva más común en el machine learning. Su objetivo es encontrar grupos naturales en los datos con base en la similitud entre instancias. Algunos ejemplos de sus aplicaciones:

- Aprendizaje de patrones de comportamiento. La agrupación puede ser utilizada para identificar las características que distinguen a un conjunto de objetos y derivar así patrones intrínsecos sobre los datos. Por ejemplo, en contextos bancarios, estos atributos podrían estar asociados con información sociodemográfica y financiera de los clientes. A partir de esta se pueden derivar grupos que permitan realizar un perfilamiento detallado de los clientes para apoyar la toma de decisiones en el otorgamiento de créditos.

- Segmentación de imágenes. Es una tarea dentro del campo de la visión artificial que permite dividir la imagen en regiones que comparten características comunes (Fig. 2). Para hacer esto se asigna una categoría a cada píxel de la imagen, lo cual se puede realizar una vez construido el modelo de agrupación. En este, píxeles parecidos serán asignados al mismo grupo.

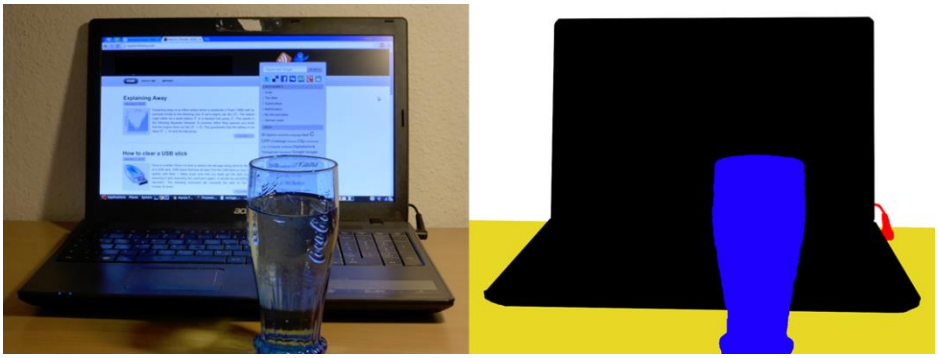

Fig. 2. Ejemplo de segmentación de imágenes. La imagen original se muestra a la izquierda. La imagen segmentada a la derecha. En esta última se puede observar que cada tipo de objeto (región) ha sido identificado con una categoría representada por un color[1]

¿Otros ejemplos? Podemos citar: identificar clientes con comportamientos de compra similares, determinar estudiantes con estilos de aprendizaje parecidos, ubicar pacientes con síntomas afines, agrupar países con economías semejantes, hallar documentos con contenidos relacionados, revelar comunidades implícitas en una red social, y muchos más.

▪ **Reducción de la dimensionalidad**.

Otra tarea común en el aprendizaje no supervisado es la reducción de la dimensionalidad, la cual puede definirse como el proceso mediante el cual se genera un espacio de entrada con un menor número de variables o características, tratando de no afectar la información que es pertinente para resolver la tarea de aprendizaje.

[1] https://commons.wikimedia.org/wiki/File:Image-segmentation-example-segmented.png

Como se observa en la Fig. 3, la disminución opera por selección o transformación. Las técnicas de selección determinan las variables más informativas del conjunto de datos, por lo que se mantiene el espacio original de representación. Esto puede lograse de varias formas, las cuales pueden trabajar en conjunto. Por ejemplo, los expertos del dominio podrían recomendar cuáles variables serían más importantes para resolver el problema. También, utilizando técnicas del análisis exploratorio es posible determinar variables poco informativas o redundantes. Por último, es posible emplear algoritmos que permitirán, de manera automática, identificar los atributos más importantes para la tarea. Estas últimas técnicas trabajan, por lo general, en contextos de clasificación.

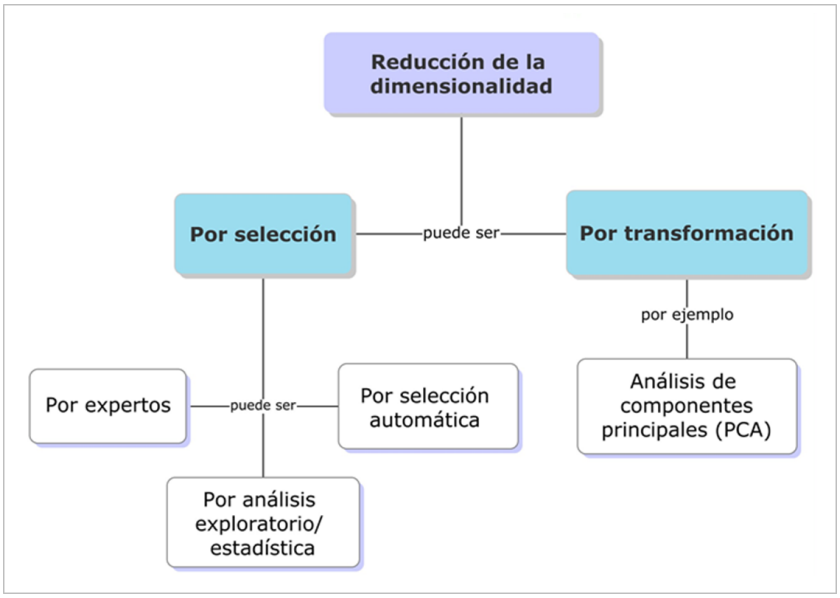

Fig. 3. Métodos para la reducción de la dimensionalidad.

El otro camino de reducción de la dimensionalidad es aplicar transformaciones a todo el conjunto de datos para generar un nuevo espacio de características que son función de las variables originales. Se intenta entonces derivar un nuevo espacio de baja dimensión, pero sin pérdida de información. En este nuevo espacio puede actuar entonces un algoritmo de aprendizaje, tanto supervisado como no supervisado, para derivar un modelo predictivo, como la clasificación, o descriptivo, como la agrupación.

Además de acelerar el aprendizaje, la reducción de la dimensionalidad es también muy útil para la visualización de datos. Por ejemplo, al reducir el número de dimensiones a dos o tres podemos proyectar el conjunto de datos de alta dimensión en un gráfico y obtener una representación visual de las relaciones entre ellos.

- **Detección de anomalías**.

En esta tarea el objetivo es detectar patrones que difieren del comportamiento esperado o normal, los cuales se conocen como anomalías, valores atípicos (*outliers*), observaciones discordantes, excepciones, sorpresas, particularidades o contaminantes en diferentes dominios de aplicación. La detección de anomalías tiene una gran aplicabilidad en problemas en diversos contextos, como: detección de productos defectuosos en líneas de producción (Fig. 4), detección de fraudes en el uso de tarjetas de crédito y en el ámbito de seguros o servicios médicos, ciberseguridad (identificación de intrusos o enlaces maliciosos), alerta de fallas en sistemas críticos y de disturbios en ecosistemas, detección de eventos raros e interesantes a partir de lecturas de sensores y revelación de patrones inusuales desde datos derivados de dispositivos médicos.

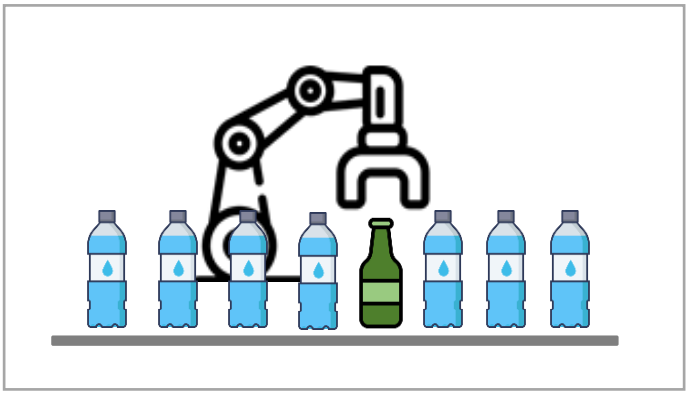

Fig. 4. Ejemplo de anomalía en una línea de producción.

En general, la estrategia a seguir para resolver esta tarea es construir un modelo normal de los datos y luego identificar los patrones que no encajan en él. Por ejemplo, si se utiliza un modelo de agrupación para resolver esta tarea, los datos normales pertenecerían a grupos grandes y densos, y los anómalos a conjuntos pequeños o dispersos o a ningún grupo.

- **Análisis de asociación**.

En el aprendizaje de reglas de asociación el objetivo es descubrir relaciones interesantes entre atributos, las cuales se derivan de patrones frecuentes encontrados en los datos. Un ejemplo icónico de aplicación de esta tarea es determinar los productos que se compran juntos con más frecuencia, cuando se quiere hacer un análisis de la cesta de compra de los clientes de un supermercado, como se muestra en la Fig. 5.

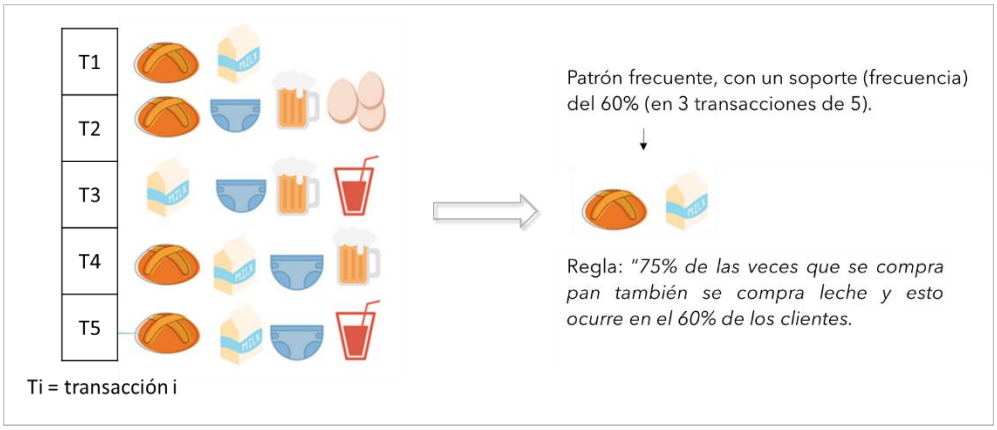

Fig. 5. Análisis de la cesta de compra.

En esta figura se pueden identificar los productos (ítems) que han comprado cinco clientes (transacciones). Se observa que un conjunto de ítems que se repite con una cierta frecuencia es [pan, leche], en este caso en el 60% de las transacciones. A partir de este se puede derivar la regla: "75% de las veces que se compra pan también se compra leche y esto ocurre en el 60% de los clientes". ¿Por qué 75% y no 100%? Porque en una compra (T2) aparece pan, pero no leche. Es decir, no podemos inferir que siempre que se compra pan se compra leche.

¿Qué utilidad tienen estos conjuntos de ítems o patrones frecuentes que se observan en los datos? Pues que definen perfiles o reglas de comportamientos en diferentes contextos. En el ejemplo, estos pueden utilizarse para la toma de decisiones sobre promociones de mercadeo, gestión de inventarios y manejo de las relaciones con los clientes. Otros contextos de aplicación del análisis de asociación son: determinar los itinerarios más seguidos por los visitantes de un sitio Web, analizar

las peticiones de servicios de laboratorio a partir de las pruebas o exámenes que frecuentemente se realizan juntas, identificar las relaciones frecuentes entre características demográficas y variables de interés en un estudio poblacional, entre otros.

Por último, ten en cuenta que al resolver un problema de aprendizaje es muy importante identificar cuál es el tipo de tarea que resulta adecuada para tratar el problema y poder decidir sobre los algoritmos de aprendizaje a aplicar. Existen algoritmos dirigidos a la tarea de agrupación, como K-medias (*K-means*). Para la reducción de la dimensionalidad podemos utilizar el Análisis de Componentes Principales (PCA) o el algoritmo T-SNE para la visualización de datos. En análisis de asociación contamos también con una diversidad de algoritmos, el más famoso es quizá Apriori. Por último, para resolver una tarea de detección de anomalías existen varios enfoques y, por lo tanto, diferentes algoritmos que podemos utilizar, como los basados en ensembles (Isolation Forest), basados en densidades (Local Outlier Factor, LOF) y basados en correlaciones (Elliptic Envelop).

Bibliografía

Geron, A. (2022). The Machine Learning Landscape. En: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (pp. 150 - 163). Second edition. O'Reilly.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.