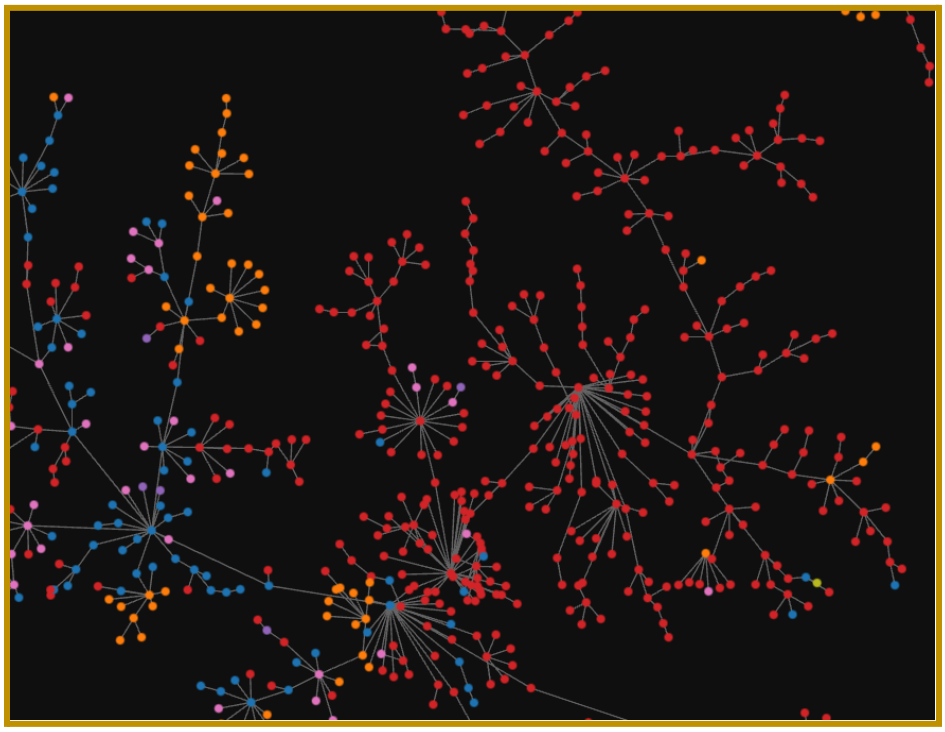
# Universidad de Guadalajara
# Centro Universitario de Ciencias Exactas e Ingenierías

*Análisis de Clustering con TMAP en base de datos*



**Asignatura:** Análisis de Algoritmos
**Sección:** D06 **Calendario:** 2025 B
**Profesor:** Lopez Arce Delgado Jorge Ernesto


Integrantes del Equipo
Nombre: Gutierrez Vazquez Axel Código:220575328
Nombre: Quintero Arreola Laura Vanessa Código: 220577347

Fecha de entrega: 5 de octubre de 2025

# Índice

Introducción
Objetivos

# Investigación Previa

## Definición de Clustering

Consiste en agrupar un conjunto de datos en grupos o clusters que compartan características similares entre sí, pero que sean diferentes de los grupos adyacentes. Mediante este método se pueden descubrir patrones o estructuras inherentes en los datos, así como identificar relaciones entre variables. Este proceso es conocido como análisis de cluster o clustering, busca clasificar los datos en función de su semejanza o distancia dentro del espacio de características, permitiendo así una organización lógica de la información.

Esta técnica resulta útil para reducir la complejidad de un conjunto de datos, ya que permite representar la información de forma más ordenada y comprensible.

## Definición de Base de datos

Una base de datos es un archivo digital organizado donde se guarda y protege una gran cantidad de información, se podría ver como una biblioteca bien estructurada para datos, que permite no solo almacenarlos, sino también administrarlos y mantenerlos seguros. Su principal función es permitir que las empresas y organizaciones manejen y guarden su información, esta información puede ser de cualquier tipo, ya sea desde datos de sus empleados hasta registros de que se ha hecho en la empresa/organización. Tomando esto en cuenta las bases de datos son fundamentales para que los datos sean accesibles y coherentes para quienes los necesitan, ya sean personas o aplicaciones.

## Reducción de dimensionalidad

La reducción de dimensionalidad es un método que se emplea para poder representar un conjunto de datos con alta dimensión y poderlos representar con una menor dimensión, al hablar de dimensiones nos referimos a las características que conforman a un dato, entonces para poder representar un conjunto de datos complejos podemos emplear la reducción de dimensionalidad para que analicemos qué características de un dato son redundantes o menos importantes pero que el valor principal de ese dato no se vea afectado y asimismo pueda reducir su dimensión.

La importancia y la utilidad de la reducción de dimensionalidad es clave para el aprendizaje automático, ya que tener demasiadas dimensiones, aunque parezca bueno, en realidad puede llegar a causar varios problemas, principalmente, consume mucho tiempo

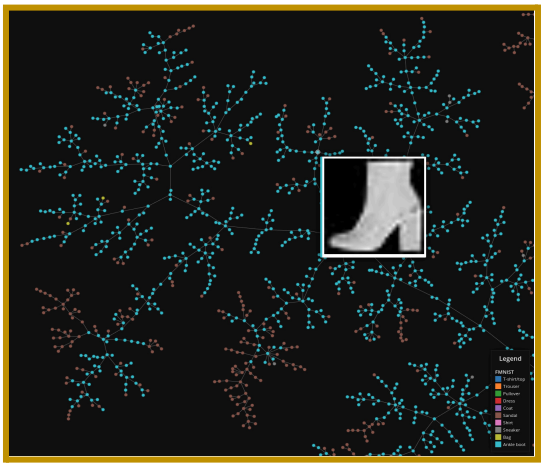
y espacio de almacenamiento, pero lo más importante es que reduce la precisión de los modelos predictivos, a este problema se le conoce como “maldición de la dimensionalidad”, esto sucede porque a medida que añadimos más y más variables, los datos se vuelven más dispersos y por ende hace que sea más difícil establecer una conexión entre todos los datos.

## TMAP

Es una librería que se utiliza principalmente para diseñar conjuntos de datos muy grandes como árboles, un ejemplo muy simple de su uso es el diseño de un gráfico. Esta herramienta permite manejar información compleja mediante estructuras como árboles, mapas o redes jerárquicas, facilitando la identificación de relaciones y conexiones entre los datos

Para hacer uso de TMAP esta necesita ser importada como una librería de la siguiente manera: import tmap.

Una vez importada, la librería permite construir representaciones visuales basadas en mapas topológicos, lo que la hace especialmente útil en campos como la bioinformática, la minería de datos, el aprendizaje automático o el análisis de similitud química. En estos contextos, TMAP ayuda a identificar agrupamientos naturales (clusters) y visualizar relaciones entre miles o millones de elementos que, de otro modo, serían imposibles de analizar visualmente.



Al representar gráficos de gran tamaño, es necesario descartar o simplificar algunas aristas con el fin de obtener un diseño más claro y fácil de interpretar. De lo contrario, el gráfico resultante podría verse saturado por un exceso de conexiones e intersecciones, dificultando la identificación de las relaciones principales entre los elementos. Esta reducción o filtrado de aristas permite resaltar las estructuras más relevantes del conjunto de datos y facilita la comprensión visual de la información representada.

El uso de TMAP no solo mejora la legibilidad de los gráficos grandes, sino que también permite identificar patrones y relaciones que serían difíciles de detectar de otra manera. Al organizar los datos de manera jerárquica y optimizar la representación de los nodos y sus conexiones, los usuarios pueden analizar conjuntos de datos masivos de forma eficiente, detectando clusters, tendencias y relaciones complejas que aportan valor a la investigación, la visualización de información o el análisis científico.

# Requisitos

**Software externo**

Python 3.8

Miniconda (opcional pero recomendado)

Editor de código o IDE (VSCode, PyCharm, Jupyter Notebook)

CSV Fashion-MNIST (fashion-mnist_train.csv)

**Librerías de Python**

tmap

networkx==2.2

pandas

numpy

scikit-learn

matplotlib

tqdm

# Desarrollo

No solo logró realizar un cluster con la librería tmap, debido a que esta era demasiado antigua y no era compatible con ninguna version del resto del entorno, se hicieron intentos por hacer que esta funcionara de varias maneras, probando crear un entorno preferencial a lo que solicitaba tmap para funcionar correctamente y aun asi no funcionaba como en la siguiente imagen se puede observar:

```
(tmap-env) C:\Users\axelg\Downloads\test>python check.py
Traceback (most recent call last):
  File "check.py", line 1, in <module>
    import tmap
ModuleNotFoundError: No module named 'tmap'
```

la imagen muestra cómo a pesar de ya tener el tmap instalado y todo lo necesario para su ejecución, este seguía sin ser detectado, finalmente tuvimos que pensar en utilizar alternativas temporales a tmap ejecución del código.

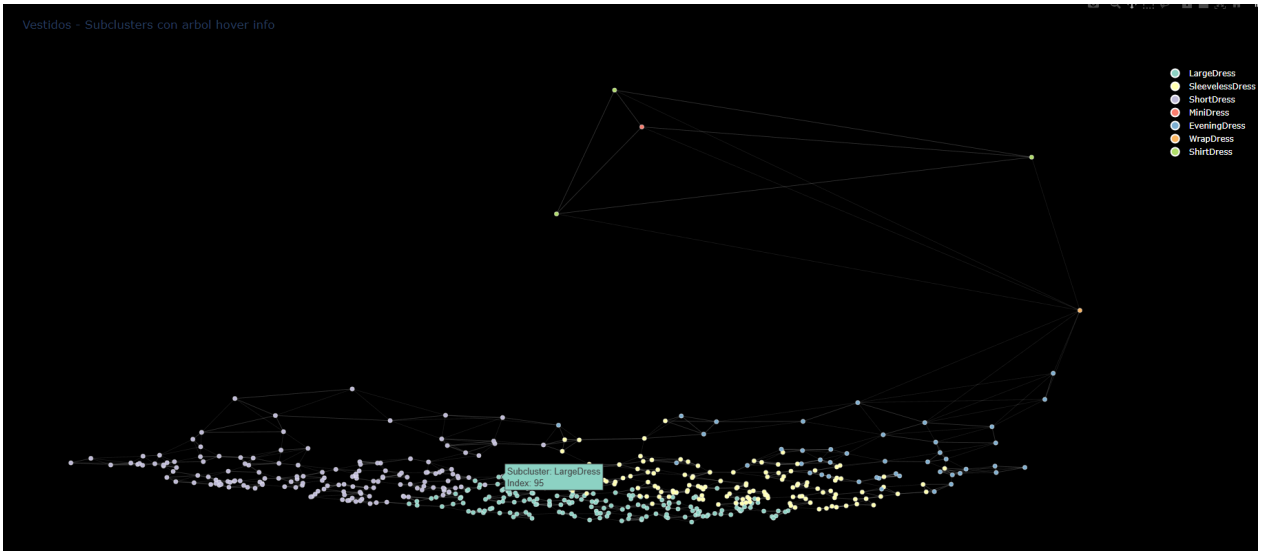
El código comienza cargando el dataset Fashion-MNIST y seleccionando un subset de filas para trabajar de forma más rápida. A continuación, filtra únicamente las filas correspondientes a vestidos, creando un conjunto de datos exclusivo para esta categoría.

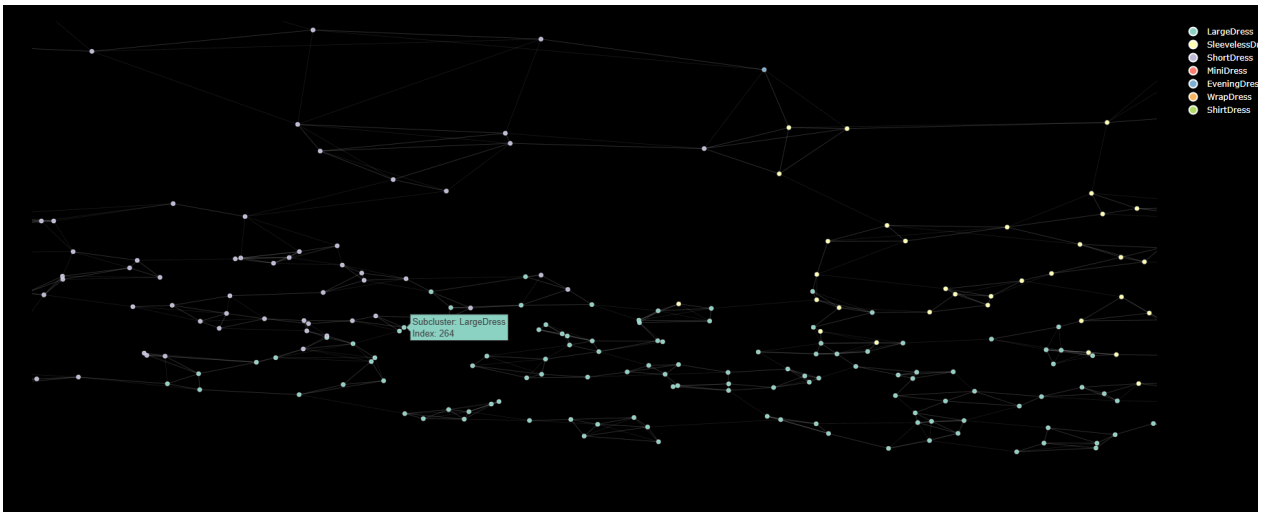
Luego normaliza las imágenes con StandardScaler y aplica PCA para reducir cada vector de 784 píxeles a coordenadas 2D. Estas coordenadas servirán para ubicar cada vestido en el gráfico interactivo.

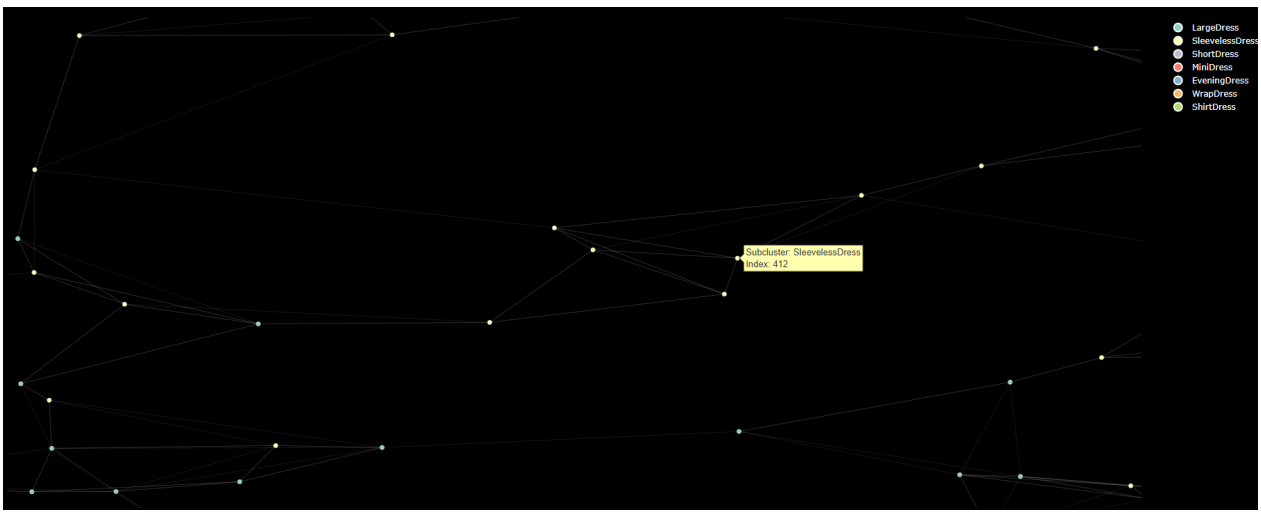
Posteriormente, se aplica KMeans para dividir los vestidos en varios subclusters, como largedress, sleevelessdress y shortdress, asignando un color distinto a cada grupo para diferenciarlos visualmente. Se generan también nodos conectados entre sí, representando la relación entre los elementos del mismo subcluster.

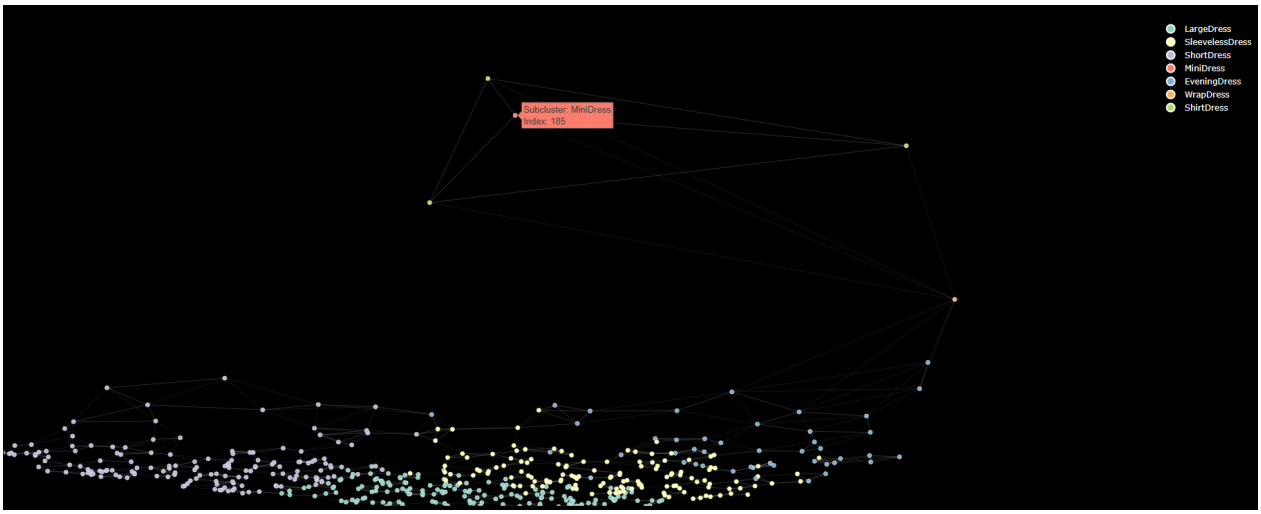
Finalmente, se crea un gráfico interactivo con Plotly donde cada nodo representa un vestido. Al poner el cursor sobre un nodo se muestra información relevante del subcluster al que pertenece y el número de elementos que contiene, reemplazando los caracteres aleatorios del hover. El resultado se guarda en un archivo HTML, permitiendo explorar los subclusters de vestidos de manera interactiva.

# Evidencia


Vestidos - Subclusters con arbol hover info
LargeDress
SleevelessDress
ShortDress
MiniDress
EveningDress
WrapDress
ShirtDress
Subcluster: LargeDress
Index: 95


LargeDress
ShortDress
MiniDress
WrapDress
ShirtDress
Subcluster: LargeDress
Index: 264


LargeDress
SleevelessDress
ShortDress
MiniDress
EveningDress
WrapDress
ShirtDress
Subcluster: SleevelessDress
Index: 412


LargeDress
SleevelessDress
ShortDress
MiniDress
EveningDress
WrapDress
ShirtDress
Subcluster: MiniDress
Index: 185

# Conclusiones

Gutierrez Vazquez Axel
Durante el desarrollo de esta actividad pudimos comprender de mejor manera cómo es que se forman los Clusters y como formarlos mediante el uso de distintas librerías, si bien tuvimos algunos problemas con el uso de la librería Tmap, logramos el objetivo mediante el uso de alternativas temporales a esta. de ese modo se

Quintero Arreola Laura Vanessa
Comprendí los aspectos teóricos de esta actividad, pero a la hora de implementar la librería de tmap no pudimos utilizarlo de forma correcta porque no es compatible con versiones actuales de python.

# Referencias


*Awan, A. A. (21 de enero de 2025). Comprender la reducción de la dimensionalidad. Datacamp.com; DataCamp.* *https://www.datacamp.com/es/tutorial/understanding-dimensionality-reduction*

*Introducción: clústeres. (30 junio de 2025). IBM. Recuperado 5 de octubre de 2025, de* *https://www.ibm.com/docs/es/was-zos/9.0.5?topic=servers-introduction-clusters*

*Ph.D, J. M., & Kavlakoglu, E. (5 de enero de 2024). Reducción de la dimensionalidad. Ibm.com.* *https://www.ibm.com/mx-es/think/topics/dimensionality-reduction*

*tmap. Documentación oficial, tmap.gdb.tools. Recuperado 5 de octubre de 2025, de* *https://tmap.gdb.tools/#support*

*Kosinski, M. (30 de septiembre de 2024). Database. Ibm.com.* *https://www.ibm.com/mx-es/think/topics/database*