Centro de Ciencias Básicas

Aprendizaje Inteligente

Actividad k-means y pca

Profesor: Dr. Francisco Javier Luna Rosas

Ingeniería en Computación Inteligente

Semestre 6° A

Integrantes:

- **Alegría Romero Dante Alejandro**
- **Aranda Gonzalez Diego Alberto**
- **Balandrán Félix Andrea Margarita**
- **Moreno Sánchez Diego Emilio**

INTRODUCCIÓN

A lo largo del semestre, hemos conocido distintos algoritmos que nos ayudan en innumerables actividades, y en esta ocasión nos encontramos con los algoritmos k-means y PCA, que son algoritmos vistos en clase.

El aprendizaje no supervisado es aquel que nos ayuda a extraer patrones y estructuras subyacentes de datos que no están etiquetados ni categorizados. Dentro de este campo, podemos encontrar un algoritmo clásico que nos muestra qué es el aprendizaje no supervisado: k-means. K-means, como se mencionó, es una técnica muy popular para el clustering. Su objetivo es dividir un conjunto de datos en k grupos, donde cada observación pertenece al grupo con el centroide más cercano. Funciona iterativamente asignando puntos de datos a los clusters más cercanos y luego actualizando los centroides en función de las observaciones asignadas. K-means es ampliamente utilizado en aplicaciones como segmentación de clientes, análisis de mercado y reconocimiento de patrones.

Por otro lado, PCA (Análisis de Componentes Principales) es una técnica de reducción de dimensionalidad utilizada para comprimir la información de un conjunto de datos en un conjunto menor de variables llamadas componentes principales. Su objetivo es encontrar las direcciones en las que los datos tienen la mayor variabilidad y proyectar los datos originales en estas direcciones. Esto permite reducir la dimensionalidad de los datos mientras se conserva la mayor cantidad posible de información. PCA es útil para visualización de datos, eliminación de características redundantes y aceleración de algoritmos de aprendizaje automático.

En la siguiente actividad, vamos a estudiar y aplicar estas dos técnicas a una tabla de Wikipedia que trata sobre la lista de Sitios del Patrimonio Mundial. Aquí, nuestro objetivo es aplicar nuestros conocimientos para utilizar el algoritmo K-Means y el PCA (Análisis de Componentes Principales) en esta tabla de datos. Así como explicaremos como es que funciona estas técnicas.

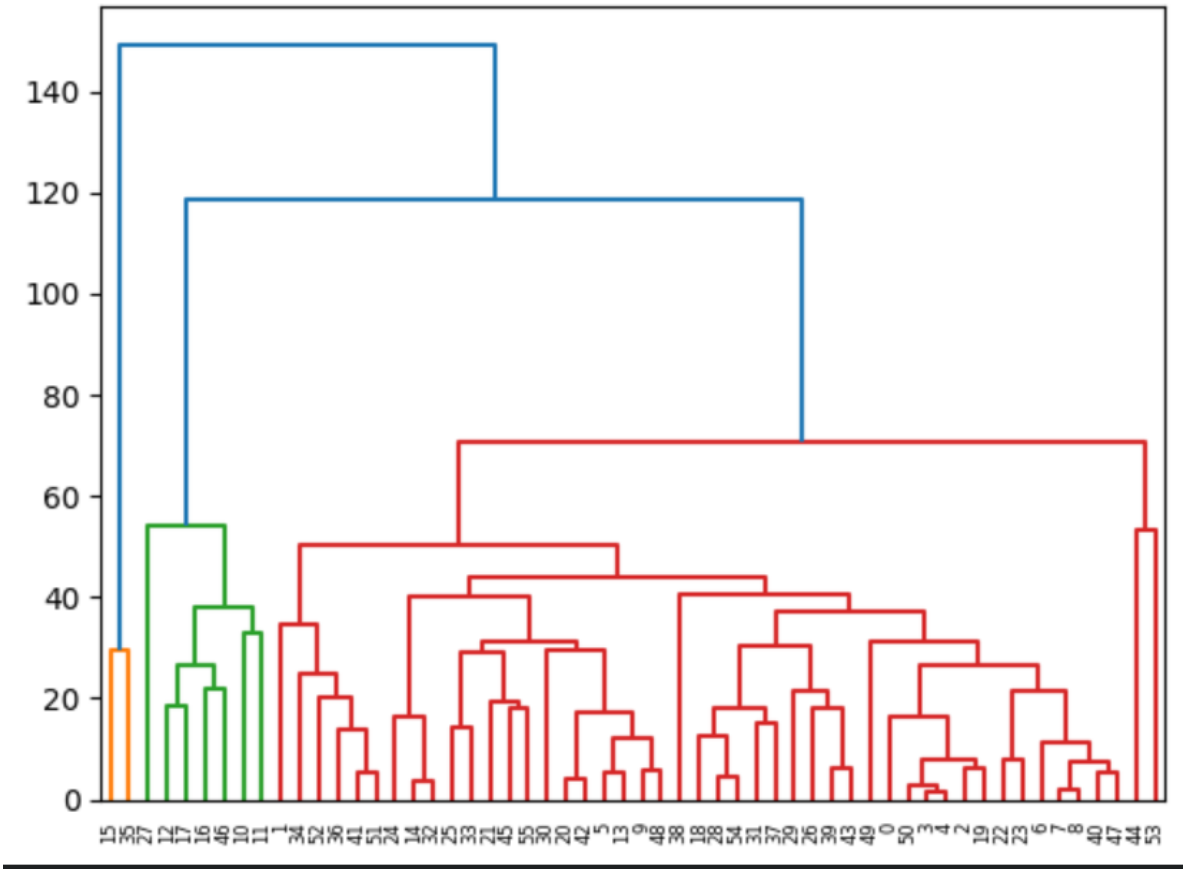
K-MEANS

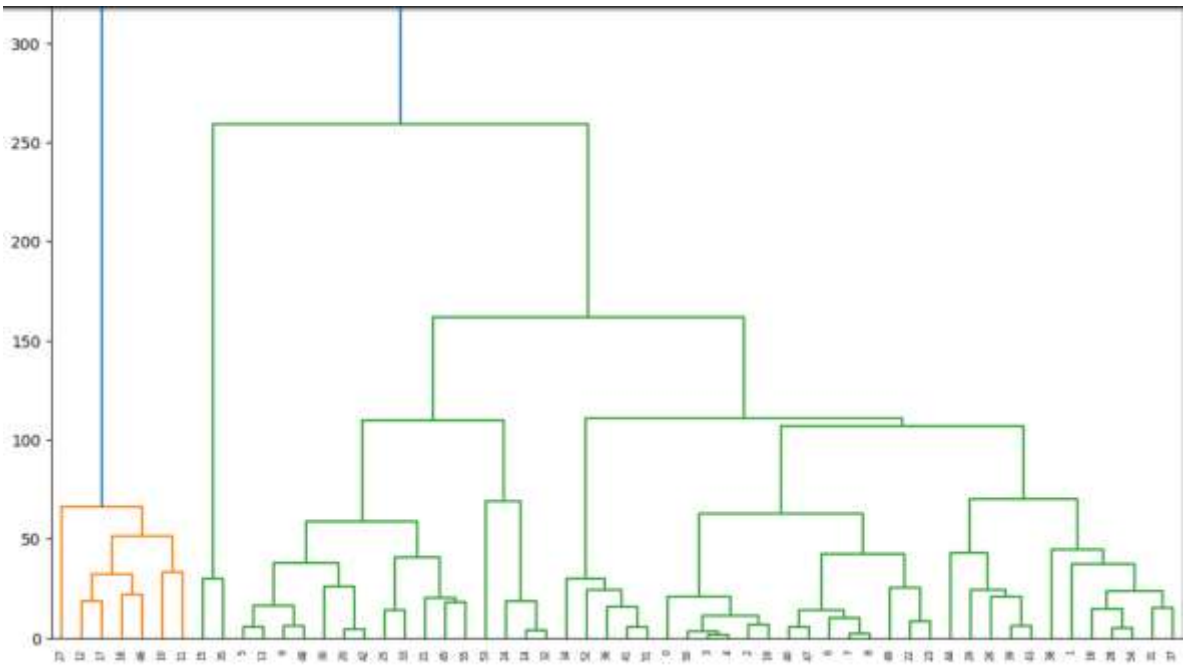
En K-Means, los datos se dividen en varios clusters, y cada cluster tiene su centroide. La técnica busca minimizar la suma de las distancias cuadradas entre cada punto de datos y el centroide de su cluster asignado.

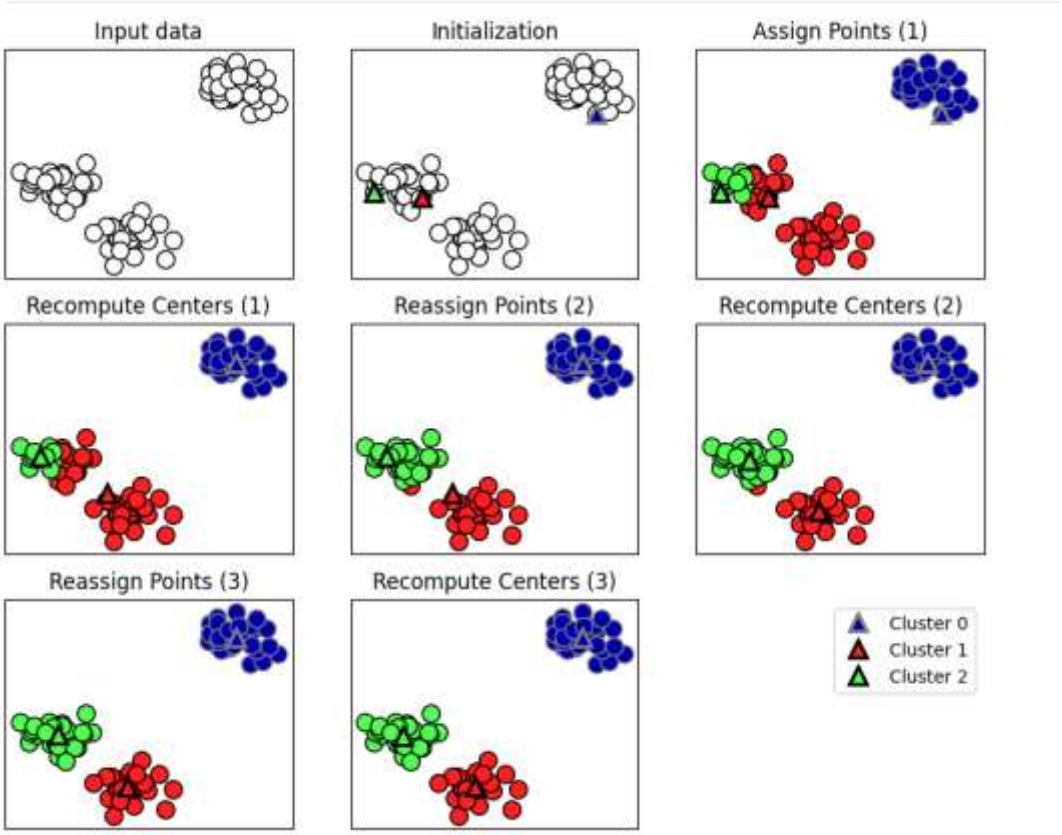
COMO FUNCIONA

1. **Inicialización de los centroides:** Primero, decidimos cuántos grupos pequeños (clusters) queremos formar. Luego, elegimos aleatoriamente algunos puntos especiales del grupo grande de datos para ser nuestros "centrales" iniciales. Estos centrales son como guías que ayudarán a organizar los datos en torno a ellos.
2. **Asignación de puntos a los clusters:** Una vez que se han seleccionado los centroides iniciales, asignamos cada punto de datos al centroide más cercano. Calculamos la distancia entre cada punto de datos y todos los centroides y asignamos el punto al cluster asociado con el centroide más cercano.
3. **Actualización de centroides:** Después de asignar todos los puntos a los clusters, recalculamos los centroides de cada cluster. Esto se hace tomando la media de todas las coordenadas de los puntos en cada cluster. Los centroides actualizados representan los nuevos centros alrededor de los cuales se agruparán los datos en la próxima iteración.
4. **Reasignación de puntos:** Una vez que se han actualizado los centroides, repetimos el proceso de asignación de puntos a los clusters utilizando los centroides actualizados. Los puntos se asignan nuevamente al centroide más cercano en función de la distancia euclidiana.
5. **Convergencia:** Este proceso de asignación y actualización de centroides se repite iterativamente hasta que se alcanza un criterio de detención, como un número máximo de iteraciones o cuando los centroides ya no cambian significativamente entre iteraciones. En cada iteración, la posición de los centroides se ajusta para minimizar la distancia entre los puntos y sus centroides asignados.
6. **Resultados finales:** Una vez que el algoritmo converge, obtenemos una partición de los datos en clusters, donde cada punto pertenece a un cluster específico. Los centroides finales representan el centro de cada cluster y se utilizan para interpretar y analizar los resultados del clustering.

A continuación, mostraré imágenes que nos explican más sobre cómo funciona el clustering







La imagen anterior es un ejemplo de como se ve los clústeres y como estos comienzan a evolucionar conforme se van transformando o evaluando.

PCA

De igual forma ya mencionamos que el pca es es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos mientras mantiene la mayor cantidad posible de información.

COMO FUNCIONA:

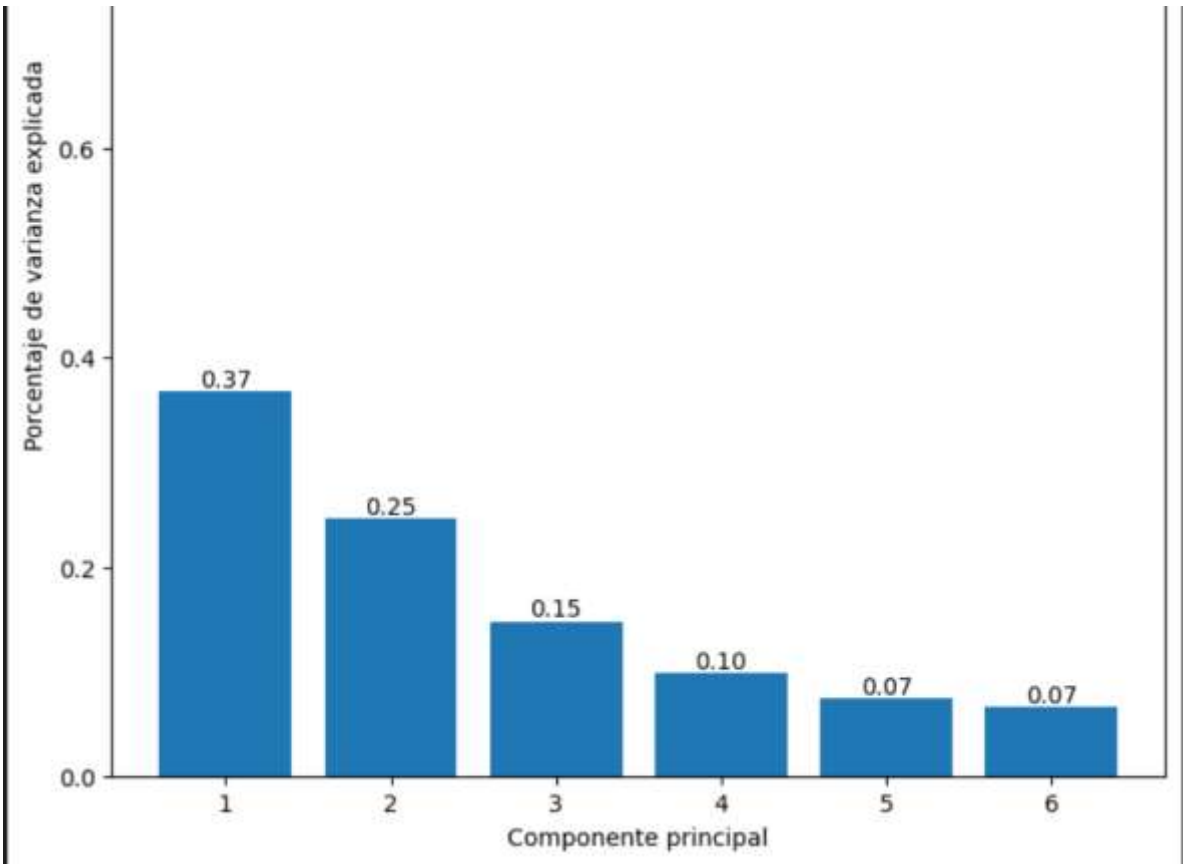
Normalización de los Datos: Antes de aplicar PCA, es importante que los datos estén normalizados, es decir, que todas las variables tengan media 0 y desviación estándar 1. Esto asegura que ninguna variable domine a las demás debido a diferencias en sus escalas.

Cálculo de Valores Propios y Vectores Propios: Se realiza un análisis de de la matriz de covarianza de los datos. Los valores propios indican la cantidad de varianza que cada componente principal explica, mientras que los vectores propios describen la dirección de estos componentes principales.

Elección de Componentes Principales: Se seleccionan los componentes principales que juntos explican una porción significativa de la varianza total de los datos. Un valor propio > 1 sugiere que el componente principal añade más varianza que cualquier una de las variables originales, siendo un criterio común para la selección inicial de componentes.

Transformación de los Datos: Finalmente, los datos se transforman en el nuevo espacio definido por los componentes principales seleccionados. Esta transformación reduce la dimensionalidad del conjunto de datos manteniendo la mayor parte de la información contenida en los datos originales.

A continuación mostrar algunas imágenes de como funciona el pca.



CONCLUSIONES:

La actividad nos permitió profundizar en el aprendizaje supervisado, una técnica clave en machine learning. Descubrimos que el aprendizaje supervisado se basa en entrenar modelos con datos etiquetados, lo que les permite aprender relaciones específicas entre las entradas y las salidas. Este proceso nos enseñó la importancia de preparar correctamente los datos y ajustar los parámetros del modelo para mejorar la precisión de las predicciones. Además, aprendimos sobre técnicas específicas como la regresión y la clasificación, así como la importancia de la validación cruzada y otras métricas de evaluación para asegurar la robustez de los modelos.