1. PCA (Análisis de Componentes Principales)

Objetivo: Reducir la dimensionalidad de un conjunto de datos conservando la mayor varianza posible.

Conceptos clave:

- **Dimensionalidad**: El número de características en un conjunto de datos. Reducir la dimensionalidad ayuda a simplificar los modelos, reducir el ruido y mejorar el rendimiento.
- **Componentes principales**: Son combinaciones lineales de las características originales que capturan la máxima varianza en los datos.
- **Varianza**: Medida de dispersión de los datos. El objetivo de PCA es identificar las direcciones (componentes) que maximizan la varianza.

Pasos en PCA:

1. **Estandarización**: Es importante escalar las características para que todas tengan la misma magnitud, ya que PCA es sensible a la escala.
2. **Covarianza**: Calcular la matriz de covarianza de los datos estandarizados. La covarianza describe cómo varían dos variables juntas.
3. **Descomposición en valores propios**: Descomponer la matriz de covarianza para encontrar los valores propios (eigenvalues) y los vectores propios (eigenvectors).
4. **Selección de componentes**: Los vectores propios con los mayores valores propios indican las direcciones de mayor varianza, y se seleccionan como componentes principales.
5. **Transformación**: Proyectar los datos originales en el nuevo espacio definido por los componentes principales.

Ventajas de PCA:

- Reducción de ruido.
- Mejora en la visualización (especialmente en 2D o 3D).
- Reducción de la complejidad computacional.

Limitaciones:

- Solo captura relaciones lineales.
- No es interpretativo (los componentes principales no tienen un significado físico claro).

2. Clustering (Agrupamiento)

Objetivo: Dividir un conjunto de datos en grupos o clústeres de objetos similares, de manera que los objetos dentro de un clúster sean más similares entre sí que con los de otros clústeres.

Tipos de Clustering:

- **Clustering Jerárquico**:
    - **Aglutinante (agglomerative)**: Comienza con cada punto como un clúster y los fusiona.
    - **Divisivo**: Comienza con todos los puntos en un solo clúster y los divide en clústeres más pequeños.
    - **Ventaja**: No necesita definir el número de clústeres previamente.
    - **Desventaja**: Computacionalmente costoso.
- **Clustering basado en centroides** (por ejemplo, K-Means):
    - Divide los datos en K clústeres según los centroides.
    - **Pasos**:
        1. Inicializar K centroides aleatorios.
        2. Asignar cada punto al centroide más cercano.
        3. Recalcular los centroides como la media de los puntos asignados.
        4. Repetir los pasos 2 y 3 hasta que los centroides no cambien.
    - **Ventajas**: Rápido y fácil de implementar.
    - **Desventajas**: Necesita definir el número de clústeres (K) de antemano. Sensible a la inicialización y a la forma de los datos.

Métodos para Evaluar la Calidad del Clustering:

- **Inercia (K-means)**: Mide la suma de distancias cuadradas de cada punto al centroide más cercano.

3. Profiling (Análisis de Perfiles)

Objetivo: Analizar y caracterizar conjuntos de datos para comprender mejor sus propiedades, detectar patrones, identificar problemas y obtener información útil.

Técnicas de Profiling:

- **Análisis exploratorio de datos (EDA)**: Usar estadísticas descriptivas y visualizaciones para explorar las características del conjunto de datos.
- **Distribución de Variables**: Analizar la distribución de cada variable para identificar anomalías, valores atípicos, sesgos, etc.
- **Correlación**: Analizar la relación entre las variables utilizando la matriz de correlación. Esto ayuda a detectar dependencias lineales.
- **Identificación de valores atípicos**: Detectar puntos que se desvían significativamente del patrón general de los datos (por ejemplo, con Boxplots).
- **Análisis de Missing Values**: Identificar y tratar los valores faltantes mediante imputación o eliminación.

Aplicaciones del Profiling:

- Mejora de la calidad de los datos.
- Identificación de patrones interesantes o desconocidos en los datos.
- Análisis de riesgos o anomalías en datos de clientes, transacciones, etc.

Resumen de Conexiones entre PCA, Clustering y Profiling

- **PCA** ayuda en la reducción de dimensionalidad antes de aplicar técnicas de clustering, ya que permite reducir el ruido y facilita que los algoritmos de agrupamiento encuentren patrones en los datos.
- **Clustering** puede ser usado después de realizar un análisis de perfiles para segmentar un conjunto de datos con características similares. La exploración inicial de los datos ayuda a entender mejor cómo se distribuyen los datos antes de realizar el agrupamiento.
- **Profiling** proporciona una base para la comprensión de los datos, permitiendo realizar un preprocesamiento adecuado antes de aplicar PCA o clustering.

1. Gráficos en PCA (Análisis de Componentes Principales)

PCA reduce la dimensionalidad de los datos, y los gráficos asociados ayudan a visualizar cómo los datos se distribuyen a lo largo de las componentes principales.

a) Gráfico de Dispersión (Scatter Plot) de las Componentes Principales

- **Objetivo**: Visualizar los datos proyectados en las primeras dos o tres componentes principales.
- **Interpretación**:
    - Cada punto en el gráfico representa una observación (fila) de los datos originales.
    - Si los puntos se agrupan en ciertas áreas, esto indica que hay agrupamientos naturales o tendencias.
    - Los puntos dispersos pueden indicar la presencia de outliers o que los datos no se ajustan a un patrón claro.

b) Gráfico de Codo (Elbow Plot)

- **Objetivo**: Ayudar a determinar cuántas componentes principales conservar.
- **Interpretación**:
    - El **codo** de la curva indica el punto donde se observa un cambio notable en la pendiente. Este es el punto donde agregar más componentes no aporta mucha varianza adicional.

c) Biplot

- **Objetivo**: Visualizar tanto las observaciones como las variables originales en el espacio reducido por PCA.
- **Interpretación**:
    - Los **puntos** muestran las observaciones proyectadas sobre las dos primeras componentes principales.
    - Las **flechas** representan las variables originales. La longitud de las flechas muestra la varianza explicada por cada variable, y la dirección de la flecha indica cómo se relaciona esa variable con las componentes principales.
    - Si las flechas están cerca unas de otras, indica que las variables están altamente correlacionadas.

2. Gráficos en Clustering (Agrupamiento)

El objetivo de los gráficos en clustering es mostrar cómo los puntos se agrupan en diferentes clústeres.

a) Gráfico de Dispersión de Clústeres (Scatter Plot)

- **Objetivo**: Visualizar cómo los datos están distribuidos en los diferentes clústeres.
- **Interpretación**:
    - Los puntos en el gráfico se colorean según el clúster al que pertenecen.
    - Puedes identificar la forma, el tamaño y la separación de los clústeres.
    - Si los clústeres están bien separados y son densos, es un buen indicativo de que el algoritmo de clustering ha funcionado correctamente.
    - Si los clústeres se solapan mucho o están dispersos, puede indicar que el número de clústeres elegido no es el adecuado, o que los datos no tienen una estructura clara.

b) Gráfico de Siluetas (Silhouette Plot)

- **Objetivo**: Medir la calidad del agrupamiento.
- **Interpretación**:
    - El valor de la **silueta** varía entre -1 y +1. Un valor cercano a +1 indica que el punto está bien asignado a su propio clúster, mientras que valores cercanos a -1 indican que el punto podría pertenecer a un clúster diferente.
    - Un gráfico de silueta muestra la puntuación de silueta para cada punto, y también el promedio de la silueta para todo el conjunto de datos.
    - Si la mayoría de los puntos tienen una silueta alta, es un buen indicio de que el clustering es adecuado.

c) Dendrograma (para Clustering Jerárquico)

- **Objetivo**: Visualizar la jerarquía de los clústeres.
- **Interpretación**:
    - El dendrograma es un árbol que muestra cómo los puntos o clústeres se agrupan a medida que se fusionan o dividen.
    - El eje vertical muestra la distancia entre los puntos o clústeres.
    - Los **ramalazos** más cercanos indican que los puntos o clústeres se agrupan rápidamente, mientras que los ramalazos más alejados indican que los puntos se agrupan a mayores distancias.
    - El **corte** horizontal del dendrograma define cuántos clústeres finales se desean (al cortar el dendrograma por encima de un cierto nivel de distancia).

3. Gráficos en Profiling (Análisis de Perfiles)

El profiling tiene como objetivo explorar y describir los datos, y los gráficos son fundamentales para entender las distribuciones y relaciones de las variables.

a) Histogramas

- **Objetivo**: Visualizar la distribución de una variable.
- **Interpretación**:
    - El **eje X** representa los intervalos de valores de la variable (bins).
    - El **eje Y** muestra la frecuencia (número de observaciones) en cada intervalo.
    - Los histogramas te permiten ver si los datos están distribuidos de manera uniforme, sesgada, o si presentan una distribución normal.
    - Las **colas largas** o picos muy pronunciados indican la presencia de outliers o distribución no uniforme.

b) Boxplots (Diagramas de Caja)

- **Objetivo**: Visualizar la distribución y los outliers de una variable.
- **Interpretación**:
    - El **cuadro** muestra el rango intercuartil (Q1 a Q3), donde se encuentra la mayoría de los datos.
    - La **línea dentro del cuadro** muestra la mediana de la distribución.
    - Los **bigotes** muestran la extensión de los datos (generalmente hasta 1.5 veces el rango intercuartil), y los puntos fuera de los bigotes se consideran outliers.
    - Los **outliers** son puntos que se desvían significativamente de la distribución central de los datos.

c) Matriz de Correlación

- **Objetivo**: Mostrar las relaciones entre diferentes variables.
- **Interpretación**:
    - Las **celdas de la matriz** muestran los coeficientes de correlación entre las variables (generalmente entre -1 y +1).
    - Un valor cercano a **+1** indica una fuerte relación positiva (cuando una variable aumenta, la otra también lo hace), mientras que un valor cercano a **-1** indica una relación negativa.

Resumen

- **PCA**: Los gráficos de dispersión, biplot y gráfico de codo son clave para evaluar cómo se distribuyen y se explican las variaciones de los datos en un espacio reducido.
- **Clustering**: Los gráficos de dispersión, siluetas y dendrogramas son útiles para entender cómo se agrupan los datos y evaluar la calidad del agrupamiento.
- **Profiling**: Los histogramas, boxplots, matrices de correlación y gráficos de valores faltantes permiten explorar y describir las características de las variables y relaciones en el conjunto de datos.