MAGNA

INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

MAGNA

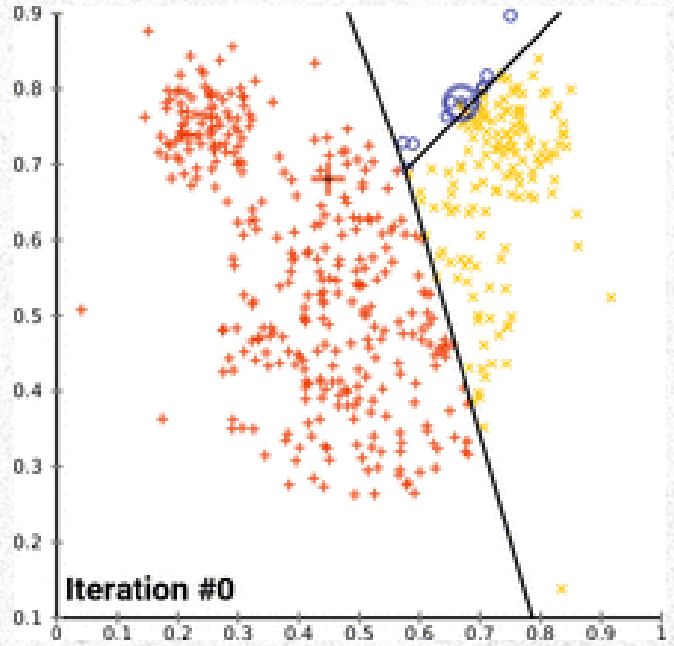
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

Algoritmo de k-means aplicado a datos de mantenimiento

Ing. Breynner Chávez

¿Qué es K-means?

El algoritmo k-means es un método de agrupamiento (clustering) que intenta organizar datos en k grupos (clusters) distintos. Cada cluster está representado por un centroide, un "centro" que define el núcleo de cada grupo. El objetivo de k-means es minimizar la variación dentro de cada grupo y maximizar la distancia entre grupos, logrando así un conjunto de clusters donde cada punto es lo más cercano posible a su centroide.



Distancia Euclidiana

La distancia euclidiana es crucial para asignar cada punto al cluster más cercano.
Es la distancia "recta" entre dos puntos, calculada así para un espacio multidimensional:

$$d(p,c) = \sqrt{\sum_{i=1}^{n} (p_i - c_i)^2}$$

donde:

- ✓ p es el punto de datos (por ejemplo, datos de rendimiento o fallos de equipos).
- ✓ c es el centroide del cluster.
- ✓ n es el número de características de los datos.

Varianza Intra-cluster y Minimización

El objetivo de k-means es minimizar la suma de la varianza intra-cluster, es decir, reducir la dispersión de los puntos respecto a sus centroides. Matemáticamente, esto se expresa como:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde:

- ✓ SSE es el "error cuadrático" o la suma de distancias cuadradas al centroide.
- ✓ k es el número de clusters.
- ✓ C_i representa los puntos en el cluster i.
- ✓ μ_i es el centroide de cada cluster.

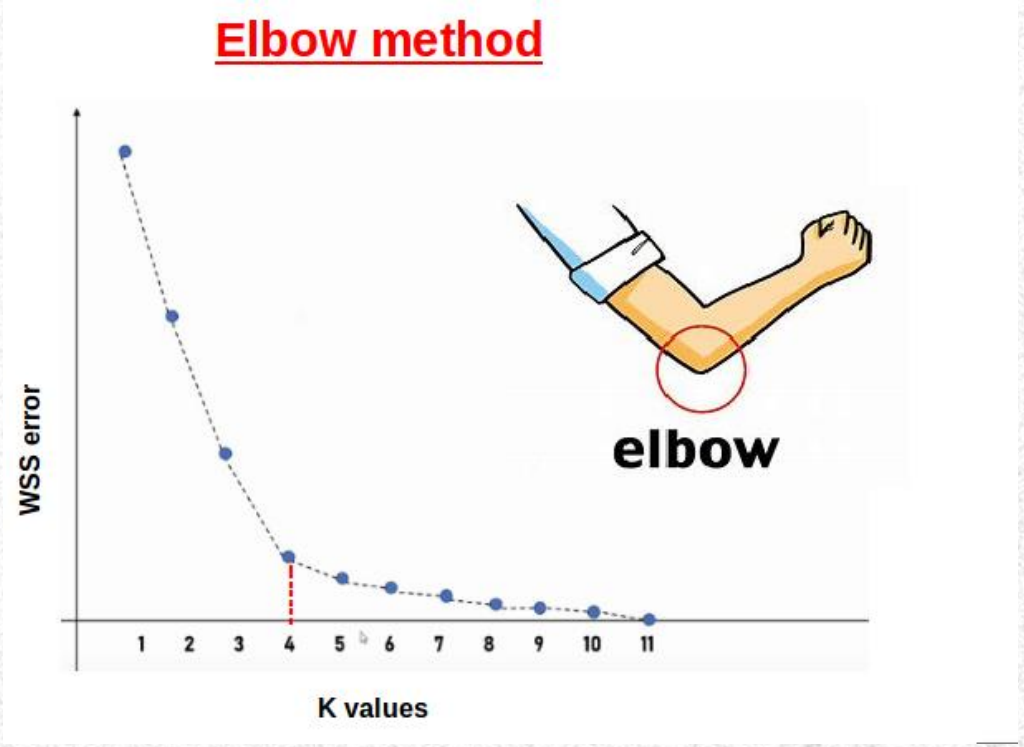
Detallando el Proceso del Algoritmo k-means Paso a Paso

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

Paso 1: Selección del Número de Clusters (k)

La elección de k es crucial y suele ser un desafío en la práctica. En la mayoría de los casos, el método del codo ayuda a encontrar el valor ideal.

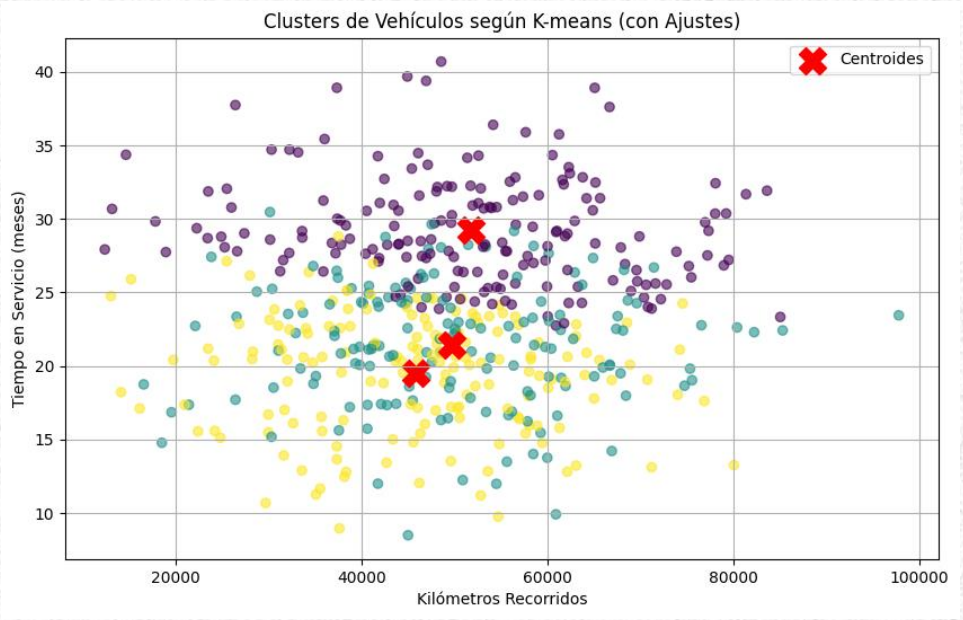
Método del Codo: Consiste en calcular el error cuadrático total para varios valores de k y observar cómo disminuye. Se elige el valor en el "codo" de la curva, donde la reducción del error empieza a ser menos significativa.



Detallando el Proceso del Algoritmo k-means Paso a Paso

Paso 2: Inicialización de Centroides

Los centroides pueden seleccionarse de manera aleatoria o mediante el método k-means++. Este último mejora la distribución inicial de los centroides, eligiendo puntos más separados para asegurar una convergencia más rápida y precisa.



Detallando el Proceso del Algoritmo k-means Paso a Paso

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

Paso 3: Asignación de Puntos a Clusters

Cada punto de datos se asigna al cluster más cercano según la distancia euclidiana a los centroides. Este proceso puede representarse matemáticamente:

$$C_{(x)} = \arg \min_{j \in \{1,\ldots,k\}} \|x - \mu_i\|^2$$

donde

- ✓ $C_{(x)}$ es el índice del cluster al cual pertenece el punto x.

Detallando el Proceso del Algoritmo k-means Paso a Paso

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

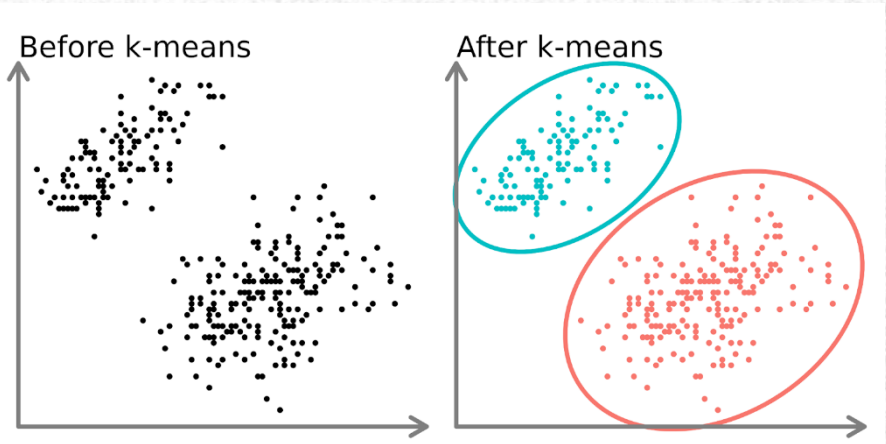
Paso 4: Recalculo de Centroides

Cada punto de datos se asigna al cluster más cercano según la distancia euclidiana
Para cada cluster, calculamos el nuevo centroide como el promedio de todos los puntos asignados. Esto puede expresarse como:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} x$$

donde:

- ✓ μ_i es el nuevo centroide del cluster i.
- ✓ C_i es el conjunto de puntos asignados al cluster i.

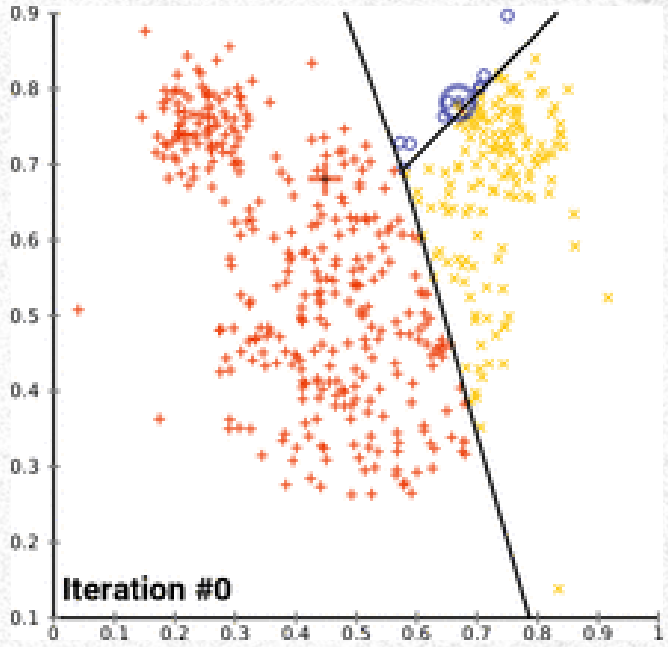


Detallando el Proceso del Algoritmo k-means Paso a Paso

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

Paso 5: Repetición y Convergencia

El algoritmo itera entre los pasos 3 y 4 hasta que los centroides no cambien de posición, lo que indica convergencia. En algunas aplicaciones, se define un número máximo de iteraciones para evitar bucles infinitos.



Aplicación Específica a Datos de Mantenimiento

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

1. ¿Por Qué Usar K-means en Mantenimiento?

- Los datos de mantenimiento suelen estar compuestos por información de rendimiento, historial de fallos, y métricas como temperatura, presión y vibración. K-means permite agrupar activos que muestran patrones similares, revelando información sobre cuándo y cómo un equipo podría requerir mantenimiento.

Aplicación Específica a Datos de Mantenimiento

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

2. Patrones Específicos Detectados con K-means:

- **Clustering de Equipos Según Vibraciones:** En el contexto de mantenimiento de maquinaria, los datos de vibración son una métrica clave, ya que un aumento en la vibración puede ser un indicio temprano de problemas como desgaste de componentes, desajustes o falta de lubricación.

- Cómo Funciona el Clustering con Datos de Vibraciones:

  - ✓ *Recopilación de Datos: Los sensores instalados en la maquinaria recopilan datos continuos de vibración. Estas lecturas pueden estar en forma de amplitud de vibración o frecuencia de vibración, medidas en diferentes condiciones de operación.*

Aplicación Específica a Datos de Mantenimiento

Clustering de Equipos Según Vibraciones

- ✓ *Recopilación de Datos: Los sensores instalados en la maquinaria recopilan datos continuos de vibración. Estas lecturas pueden estar en forma de amplitud de vibración o frecuencia de vibración, medidas en diferentes condiciones de operación.*

  - ❑ *Los datos de vibración se agrupan usando k-means, dividiéndolos en clusters.*
  - ❑ *Para este caso, podemos definir k=3, indicando tres grupos:*
    1. *Cluster de Bajo Riesgo: Equipos con vibraciones normales o dentro del rango esperado, indicando un buen estado.*
    2. *Cluster de Riesgo Medio: Equipos con vibraciones superiores al promedio, sugiriendo que puede ser necesario revisar ciertos componentes en un futuro próximo.*
    3. *Cluster de Alto Riesgo: Equipos con altos niveles de vibración, indicando posibles fallos inminentes o deterioro avanzado.*

Aplicación Específica a Datos de Mantenimiento

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

Clustering de Equipos Según Vibraciones

- ✓ *Resultados del Clustering:*

  - ❑ Cada equipo se asigna a uno de estos clusters en función de su patrón de vibración. Los técnicos pueden priorizar la inspección de los equipos del cluster de alto riesgo y programar reparaciones preventivas.

  - ❑ Además, el análisis de clusters permite entender patrones de desgaste y correlacionarlos con otros factores, como la antigüedad del equipo o el tipo de uso, lo que ayuda a optimizar el mantenimiento y aumentar la vida útil de la maquinaria.

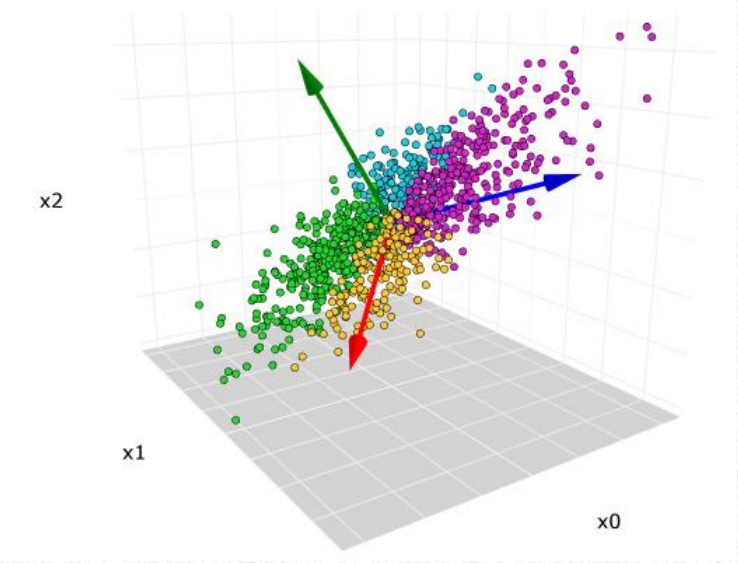
MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

Tema 2: Reducción de dimensionalidad con PCA en conjuntos de datos complejos

¿Qué es PCA?

El Análisis de Componentes Principales (PCA) es una técnica de análisis estadístico que transforma un conjunto de observaciones de variables correlacionadas en un conjunto de valores de variables no correlacionadas llamadas componentes principales.
Estos componentes se ordenan de tal manera que el primer componente captura la mayor parte de la varianza de los datos, el segundo componente captura la segunda mayor parte de la varianza, y así sucesivamente.

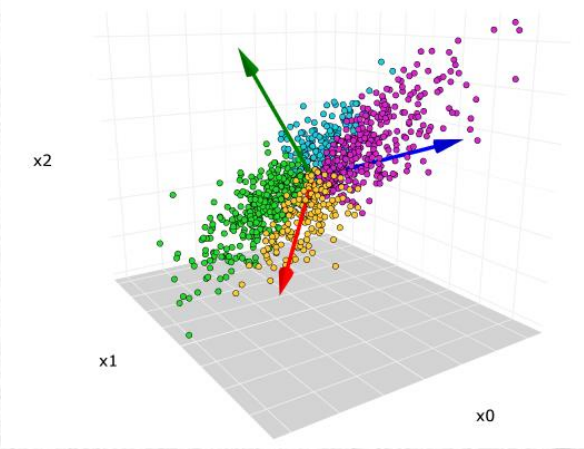
Ciencia de Datos: Para reducir la dimensionalidad de conjuntos de datos masivos y mejorar la eficiencia del procesamiento, es una herramienta poderosa y versátil para simplificar y entender datos complejos. Desde la reducción de dimensionalidad hasta la mejora en la visualización



Importancia de PCA

1. Reducción de dimensionalidad: En muchos campos, especialmente en el análisis de datos, se manejan conjuntos de datos con muchas dimensiones (variables). Esto puede llevar a problemas de "la maldición de la dimensionalidad", donde el rendimiento de los modelos de aprendizaje automático puede verse afectado negativamente debido a la complejidad y el ruido en los datos.

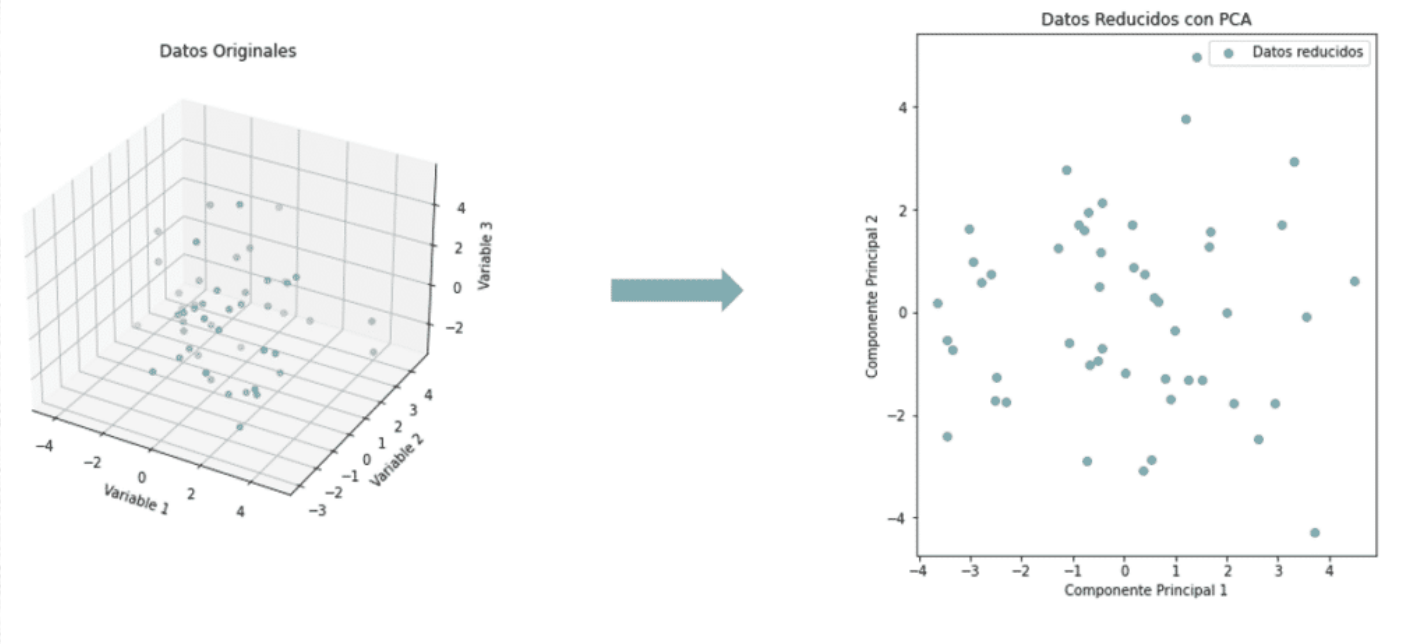
2. Visualización de datos: PCA permite visualizar datos multidimensionales en dos o tres dimensiones, lo que facilita la identificación de patrones y relaciones.



Importancia de PCA

3. Eliminación de ruido: Al reducir el número de dimensiones, PCA puede ayudar a eliminar la variabilidad que no es relevante, haciendo que los patrones en los datos sean más claros.

4. Facilitación de algoritmos de aprendizaje automático: Los algoritmos suelen funcionar mejor con un número reducido de características, ya que se pueden reducir las interacciones no deseadas y mejorar el rendimiento.



Cómo funciona PCA

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

1. **Estandarización:** Es fundamental que las variables tengan la misma escala. Esto se realiza restando la media y dividiendo por la desviación estándar de cada variable. La estandarización asegura que las variables que tienen diferentes unidades no dominen la variabilidad.

$$Z_i = \frac{X_i - \mu}{\sigma}$$

donde:

- ✓ Z_i es el valor estandarizado.
- ✓ X_i es el valor original.
- ✓ μ es la media.
- ✓ σ es la desviación estándar.

Cómo funciona PCA

2. **Cálculo de la matriz de covarianza:** Esta matriz describe la variación y la relación entre las variables. Se calcula como:

$$Cov(X,Y) = \frac{1}{n-1}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

donde:

- ✓ Cov(X,Y) es la covarianza entre dos variables.
- ✓ n es el número total de observaciones.

Cómo funciona PCA

3. **Cálculo de la matriz de covarianza:** Cálculo de valores y vectores propios: Los valores propios indican la cantidad de varianza capturada por cada componente, y los vectores propios indican las direcciones de esos componentes.

   Se resuelve el polinomio característico de la matriz de covarianza:

$$|Cov - \lambda I| = 0$$

   donde:

   - ✓ λ es el valor propio.
   - ✓ I es la matriz identidad.

Cómo funciona PCA

4. **Selección de componentes:** Se ordenan los valores propios de mayor a menor. Se seleccionan los primeros k valores propios que capturan una proporción significativa de la varianza total (usualmente se busca capturar el 80-90% de la varianza).
5. **Proyección de los datos:** Finalmente, los datos originales se proyectan sobre los vectores propios seleccionados. Esto se realiza multiplicando la matriz de datos estandarizados por la matriz de vectores propios seleccionados.

Fórmula: $Y=XW$

donde:

- ✓ Y es el nuevo conjunto de datos.
- ✓ X es la matriz de datos estandarizados.
- ✓ W es la matriz de vectores propios seleccionados.

Ejemplos Prácticos

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

1. Reducción de datos de sensores de una turbina

- **Contexto:** Imagina que tienes una turbina con varios sensores que registran parámetros como temperatura, presión, vibración y velocidad. Estos datos se recopilan continuamente y pueden ser utilizados para detectar fallas antes de que ocurran.

- **Teoría:**
  - **Datos multidimensionales:** Los sensores producen múltiples lecturas al mismo tiempo, lo que genera un conjunto de datos con cuatro dimensiones (una por cada parámetro). Esta multidimensionalidad puede dificultar la detección de patrones o anomalías.

Ejemplos Prácticos

1. Reducción de datos de sensores de una turbina

- **Teoría:**
  - **Aplicación de PCA:**
    1. **Estandarización:** Se escalan las lecturas de cada sensor. Supongamos que la temperatura oscila entre 20 y 100 grados Celsius, mientras que la presión está entre 1 y 10 bar. Sin estandarización, la presión dominaría la matriz de covarianza debido a su mayor rango de valores.
    2. **Cálculo de la matriz de covarianza:** Se obtiene una matriz de covarianza que muestra cómo varían juntas las distintas medidas. Por ejemplo, podría haber una alta covarianza entre temperatura y vibración, lo que indica que a medida que la temperatura aumenta, también lo hace la vibración.

Ejemplos Prácticos

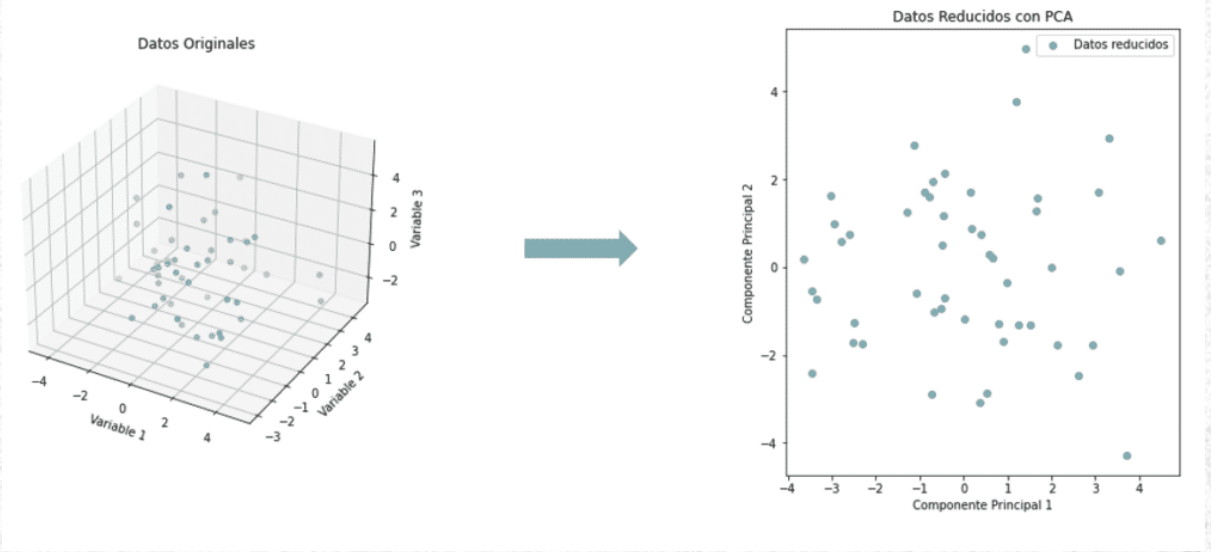
1. Reducción de datos de sensores de una turbina

- **Teoría:**
  - **Aplicación de PCA:**
    3. **Cálculo de valores y vectores propios:** Al calcular los valores y vectores propios, supongamos que encontramos que el primer valor propio tiene un alto valor, indicando que este componente explica la mayor parte de la varianza en los datos.
    4. **Selección de componentes:** Si el primer componente explica el 70% de la varianza y el segundo el 20%, podemos decidir retener estos dos componentes para análisis futuros.
    5. **Proyección:** Proyectamos los datos originales en el espacio definido por los dos componentes seleccionados. Ahora, en lugar de trabajar con cuatro dimensiones, trabajamos con un nuevo conjunto de datos bidimensional.

Ejemplos Prácticos

1. Reducción de datos de sensores de una turbina

- **Beneficio:** Esta reducción permite a los ingenieros visualizar rápidamente los datos en un gráfico de dispersión, facilitando la identificación de patrones o puntos que podrían representar un funcionamiento anómalo. Por ejemplo, si los puntos se agrupan de manera inusual en el gráfico, esto puede indicar un problema potencial que requiere atención.



Ejemplos Prácticos

2.Optimización de datos históricos de mantenimiento

- **Contexto:** Las empresas suelen recopilar grandes volúmenes de datos sobre el mantenimiento de equipos. Estos datos pueden incluir el tiempo de funcionamiento, las fechas de mantenimiento, los tipos de fallas, y las condiciones operativas, lo que genera un conjunto de datos muy amplio y complejo.

- **Teoría:**
  - **Datos complejos:** Cada variable en el conjunto de datos puede contener información útil sobre el rendimiento del equipo. Sin embargo, la cantidad de datos puede dificultar el análisis efectivo, haciendo que sea complicado identificar tendencias o factores que afecten la durabilidad de los equipos.

Ejemplos Prácticos

2.Optimización de datos históricos de mantenimiento

- **Teoría:**
  - **Aplicación de PCA:**
    1. **Estandarización:** Las diferentes variables se estandarizan para que tengan igual peso en el análisis. Por ejemplo, los tiempos de funcionamiento pueden estar en horas, mientras que las fechas de mantenimiento están en días.
    2. **Cálculo de la matriz de covarianza:** Se calcula la matriz de covarianza para entender cómo se relacionan las variables. Esto puede revelar que ciertas variables están altamente correlacionadas, como el tiempo de funcionamiento y la frecuencia de mantenimiento.

Ejemplos Prácticos

2.Optimización de datos históricos de mantenimiento

- **Teoría:**
  - **Aplicación de PCA:**
    3. **Cálculo de valores y vectores propios:** Se obtienen los valores y vectores propios, mostrando qué combinaciones de variables capturan más variabilidad en el conjunto de datos.
    4. **Selección de componentes:** Supongamos que seleccionamos los primeros tres componentes, que juntos explican el 85% de la varianza total. Esto significa que hemos logrado reducir significativamente la complejidad del conjunto de datos.
    5. **Proyección:** Proyectamos el conjunto de datos original en estos tres componentes. Ahora tenemos un conjunto de datos que es más fácil de manejar y que aún retiene la mayor parte de la información relevante.

Ejemplos Prácticos

2.Optimización de datos históricos de mantenimiento

- **Beneficio:** Con el conjunto de datos reducido, los analistas pueden identificar más fácilmente los factores clave que influyen en la durabilidad del equipo. Por ejemplo, pueden observar que las máquinas que operan a un nivel de carga más alto tienden a requerir mantenimiento con mayor frecuencia.

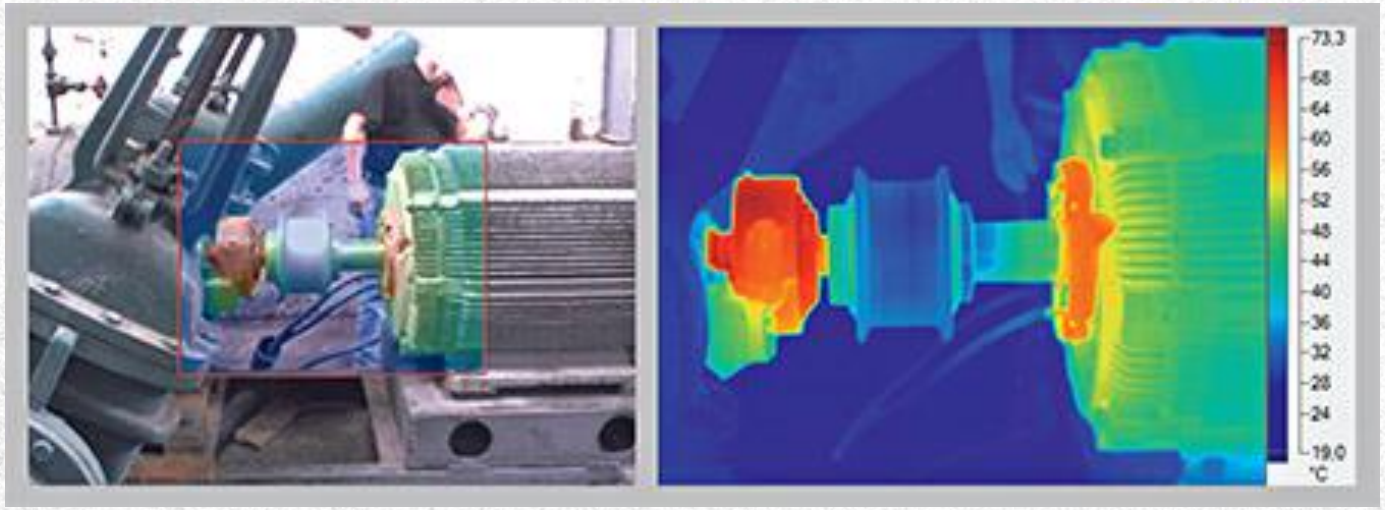
- Esta información puede ser valiosa para optimizar los programas de mantenimiento y prever problemas antes de que ocurran, mejorando la eficiencia operativa.

Tema 3: Visualización avanzada de clustering y análisis exploratorio

La visualización avanzada es esencial en el análisis de datos de mantenimiento, permitiendo a los ingenieros y técnicos observar patrones y relaciones complejas de manera intuitiva. En particular, las visualizaciones de clustering ayudan a interpretar los resultados de algoritmos no supervisados como k-means y PCA, mostrando cómo se agrupan los datos y destacando puntos anómalos.

1. Mapa de Calor para Monitorear Temperaturas en Equipos

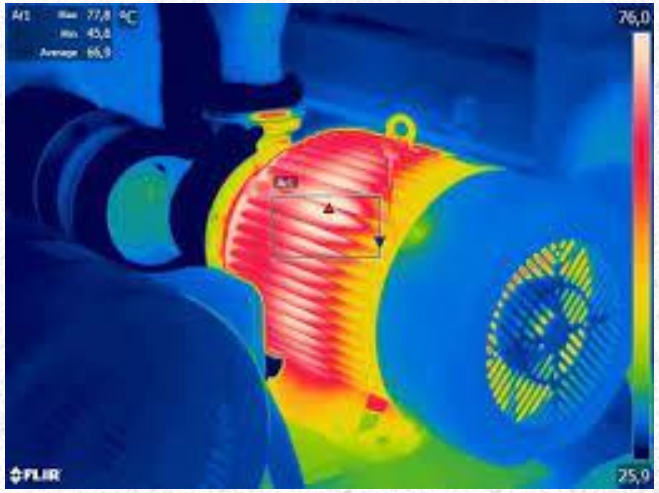
Los mapas de calor son herramientas visuales que utilizan colores para representar datos, y son especialmente útiles en contextos donde se necesita interpretar rápidamente patrones o tendencias en grandes volúmenes de información. En el caso del monitoreo de temperaturas en equipos, un mapa de calor puede revelar áreas problemáticas en una planta.



¿Por qué son útiles los mapas de calor?

- **Visualización Intuitiva:** Permiten a los usuarios identificar fácilmente áreas de interés sin necesidad de analizar datos numéricos complejos.

- **Detección de Anomalías:** Los cambios de color pueden indicar condiciones fuera de lo normal, facilitando la identificación de posibles problemas antes de que se conviertan en fallas.

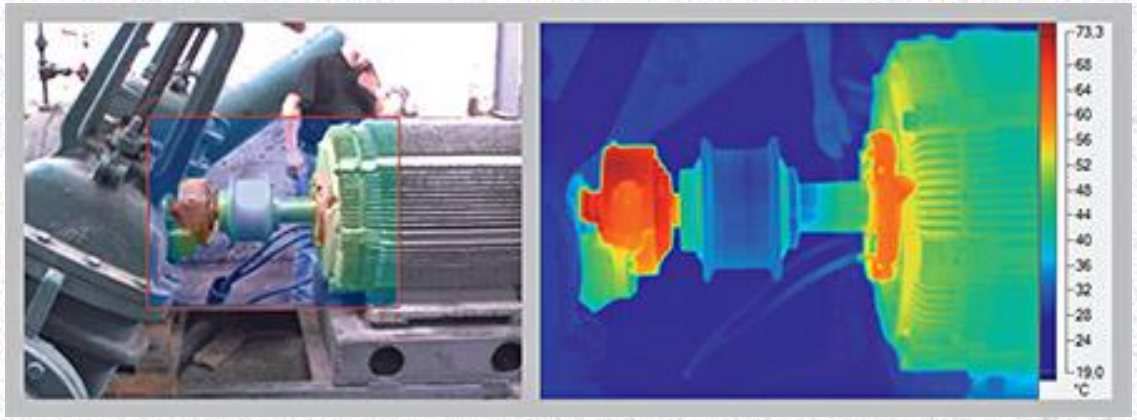


Datos Necesarios:

- **Temperatura de los Equipos:** Datos recolectados a través de sensores. Por ejemplo, cada sensor puede registrar la temperatura cada 10 minutos.

- **Ubicación de los Equipos:** Es importante tener un plano de la planta que muestre la ubicación de cada equipo para poder mapear correctamente los datos.

- **Tiempo de Monitoreo:** Se requiere una serie temporal para analizar cómo cambian las temperaturas a lo largo del tiempo.



Proceso para Crear un Mapa de Calor:

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

- **Recolección de Datos:** Recopilar datos de temperatura de todos los equipos en intervalos regulares. Por ejemplo, puedes usar un sensor que envíe datos cada 10 minutos durante un periodo de una semana.
- **Preprocesamiento de Datos:** Limpiar los datos para eliminar outliers o errores de lectura. Esto podría incluir la eliminación de datos que son físicamente imposibles (por ejemplo, temperaturas extremadamente bajas o altas).
- **Agrupación de Datos:** Agrupar los datos en matrices, donde las filas representan las distintas ubicaciones y las columnas representan los distintos tiempos de lectura.
- **Generación del Mapa de Calor:** Utilizar herramientas de visualización (como matplotlib en Python) para crear el mapa de calor. Cada celda en el mapa representará la temperatura en un momento específico para un equipo específico.

Ejemplo Práctico

Supongamos que estás monitoreando las temperaturas de cinco equipos en diferentes ubicaciones de una planta durante una semana. Los datos recolectados son los siguientes:

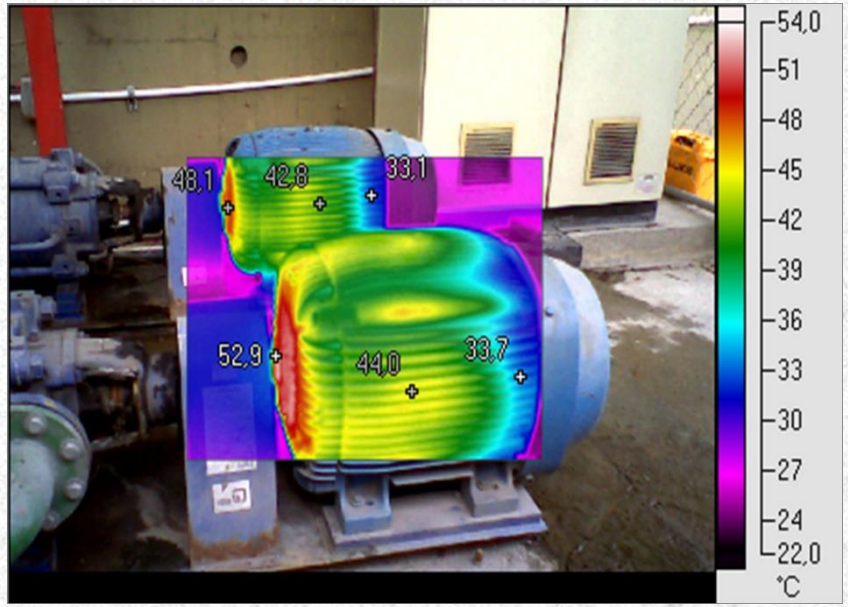
Tiempo	Equipo 1	Equipo 2	Equipo 3	Equipo 4	Equipo 5
Día 1 08:00	70°C	72°C	68°C	75°C	71°C
Día 1 12:00	75°C	76°C	70°C	80°C	73°C
Día 1 16:00	78°C	74°C	72°C	82°C	76°C
Día 2 08:00	71°C	73°C	67°C	76°C	70°C
Día 2 12:00	76°C	78°C	71°C	81°C	74°C

Con estos datos, puedes generar un mapa de calor donde los colores representan las temperaturas. Un color rojo puede indicar temperaturas por encima de 80°C, lo que podría ser una señal de advertencia para el Equipo 4.

Ejemplo Práctico

Interpretación del Mapa de Calor:

- Si en el mapa de calor observas que el Equipo 4 muestra consistentemente colores rojos, esto podría indicar un problema de sobrecalentamiento.
- Comparando las temperaturas entre equipos, podrías notar que el Equipo 3 tiene temperaturas consistentemente más bajas, lo que podría indicar un mejor funcionamiento o una mala medición.



2.Gráficos de Dispersión para Clusters de Fallas

Los gráficos de dispersión son herramientas efectivas para visualizar la relación entre dos variables numéricas. En el contexto del análisis de fallas, los gráficos de dispersión permiten visualizar cómo diferentes equipos se agrupan según características específicas, lo que facilita la identificación de patrones.

Ventajas de los Gráficos de Dispersión:

- **Identificación de Tendencias:** Puedes observar fácilmente la relación entre dos variables, como la vibración y el consumo energético.

- **Detección de Outliers:** Los puntos que se alejan del resto pueden indicar anomalías o problemas en el funcionamiento de los equipos.

Datos Necesarios:

- **Niveles de Vibración:** Medidas de vibración de los equipos, que pueden ser recolectadas por sensores.
- **Consumo Energético:** Datos de consumo energético de los mismos equipos durante el mismo periodo de tiempo.
- **Identificación de Equipos:** Cada punto en el gráfico debe identificarse con el nombre o número del equipo correspondiente.

Proceso para Crear un Gráfico de Dispersión:

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

1. **Recolección de Datos:** Obtener datos de vibración y consumo energético en el mismo intervalo de tiempo.
2. **Normalización de Datos:** Es posible que los rangos de las dos variables sean diferentes; la normalización puede ayudar a que la visualización sea más efectiva.
3. **Aplicación de k-means:** Agrupar los datos utilizando el algoritmo k-means. Esto te permitirá identificar clusters basados en las características de vibración y consumo energético.
4. **Generación del Gráfico de Dispersión:** Utilizar herramientas como matplotlib o seaborn en Python para crear el gráfico de dispersión.

Ejemplo Práctico

Supongamos que estás monitoreando las temperaturas de cinco equipos en diferentes ubicaciones de una planta durante una semana. Los datos recolectados son los siguientes:

Máquina	Vibración (mm/s)	Consumo Energético (kW)
A	1.5	10
B	2.0	12
C	1.8	11
D	3.5	15
E	2.2	13

Ejemplo Práctico

MAGNA
INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL

1. **Visualización:**
   - Al graficar estos datos, el eje X representa la vibración y el eje Y representa el consumo energético.

2. **Clusters:**
   - Al aplicar el algoritmo k-means, podrías identificar que las máquinas A, B y C se agrupan juntas, mientras que las máquinas D y E forman un grupo separado.

3. **Interpretación:**
   - Si las máquinas D y E están agrupadas, puede ser un indicativo de que tienen un comportamiento similar que merece una inspección más profunda. Además, si hay puntos que se alejan del resto (outliers), esto podría indicar equipos que están fallando o funcionan de manera ineficiente.

MAGNA

INSTITUCIÓN DE ESPECIALIZACIÓN PROFESIONAL