

**INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO**

**LETICIA VEGA MORALES**

**EVA CONSUELO MORAILA LOPEZ**

**PROFESOR: GADDIEL DESIRENA LÓPEZ**

**24 DE SEPTIEMBRE 2025**

**Clustering de datos**

**Cáncer Cervicouterino**

El Cáncer la principal causa de muerte en el mundo, donde se tienen estadística que lo más comunes son mama, pulmón, colon, recto y próstata

Entre las causas principales del cáncer son el fumar, índice alto de masa corporal, el consumo de alcohol, falta de actividad física y comer poco frutas y verduras, no podemos dejar de lado la contaminación del aire

Al hablar del cáncer cervicouterino es tratable si es detectado a tiempo, pero se dice que es una de las causas de mortalidad en mujeres.

**Objetivo de este proyecto**: Trabajar con un dataset con información de factores de riesgo relacionado con el cáncer cervicouterino, realizando un análisis exploratorio de los datos con técnicas de clustering.

**Desarrollo:**

Este análisis fue explorar los datos del archivo *sobar-72.csv* y descubrir patrones ocultos mediante técnicas de clustering. El clustering nos permite identificar grupos naturales dentro de los datos, sin necesidad de que alguien los defina previamente. Es decir, los datos “hablan” y nos muestran qué registros se parecen entre sí y cómo se diferencian de los demás.

**1. Exploración inicial de los datos**

El archivo contiene:

* 72 registros.
* 20 variables numéricas.

Hallazgo positivo: no había valores faltantes en las variables numéricas. Esto significa que pudimos trabajar con la totalidad del dataset, sin necesidad de imputar o descartar información. Esto da más confianza en que los resultados obtenidos reflejan la realidad completa de los datos.

**2. Extracción de estadísticas básicas**

Antes de aplicar algoritmos de clustering, exploramos las estadísticas principales de las variables. Algunos patrones observados:

La mayoría de las variables tienen medias en torno a la mitad de la escala (lo cual sugiere que no hay sesgos extremos).

Existen variables con alta dispersión, lo que indica que no todos los individuos perciben o experimentan las mismas cosas.

Algunas variables relacionadas, como apoyo social y empoderamiento, muestran tendencias a moverse juntas.

Esto nos sugiere que algunos factores humanos están relacionados y podrían formar parte de un mismo perfil.

**3. Reducción de la complejidad con PCA**

Como las 20 variables pueden ser difíciles de visualizar al mismo tiempo, aplicamos un Análisis de Componentes Principales (PCA).

**Los resultados fueron:**

Primera componente: explica aproximadamente 29% de la variabilidad total.

Segunda componente: explica alrededor de 18%.

En conjunto, casi el 47% de la variabilidad se puede observar en un gráfico bidimensional.

PCA (2 comps)

Varianza explicada:

PC1 ≈ 29.1%

PC2 ≈ 18.6%

PC1+PC2 ≈ 47.7% suficiente para una vista 2D útil pero no explica toda la varianza.

Esto nos permitió visualizar los datos de manera más sencilla y detectar posibles agrupaciones en un espacio reducido.

**4. Correlaciones y relaciones entre variables**

Al revisar la matriz de correlación, encontramos que:

* Las variables relacionadas con empoderamiento y apoyo social tienden a estar correlacionadas. Esto sugiere que las personas que sienten mayor empoderamiento también suelen percibir mayor apoyo en su entorno.
* Hay variables que son más independientes, lo que indica que no todos los factores se mueven juntos.
* No todos los individuos tienen el mismo equilibrio entre motivación, empoderamiento y apoyo social.

**5. Determinación del número de clusters**

Para descubrir cuántos grupos forman los datos, usamos dos criterios diferentes:

**Método del codo:** observamos la “curva de mejora” de los modelos al aumentar el número de clusters. El punto donde la curva deja de mejorar drásticamente se considera el número óptimo.

**Método del gradiente (curvatura):** mide dónde ocurre el mayor “cambio brusco” en la reducción de la inercia.

**Resultados:**

* KMeans: ambos criterios sugirieron 3 clusters como número óptimo.
* Clustering jerárquico: también coincidió en 3 clusters como la mejor partición.

La coincidencia entre ambos métodos refuerza la solidez de la conclusión: los datos se agrupan naturalmente en tres grandes perfiles.

**6. Perfiles de los clusters**

Al analizar las medias de cada grupo, interpretamos lo siguiente:

**Cluster 1**: El grupo motivado y empoderado, representa a los individuos más fortalecidos y listos para enfrentar retos

* Altos niveles de motivación.
* Buen conocimiento sobre empoderamiento.
* Consistencia en actitudes positivas.
* Buena percepción de apoyo social.

**Cluster 2:** El grupo intermedio, podrían beneficiarse de estrategias selectivas de refuerzo, dependiendo del área que necesiten mejorar

* Mezcla de fortalezas y debilidades.
* Destacan en algunas áreas, pero muestran carencias en otras.

**Cluster 3:** El grupo vulnerable, son quienes necesitan mayor acompañamiento y apoyo cercano para progresar

* Bajos niveles de empoderamiento y motivación.
* Menor percepción de apoyo social.
* Actitudes menos consistentes.

**7. Insights principales**

* Se utilizaron los 72 registros originales.
* Existen factores comunes empoderamiento y apoyo social, que tienden a reforzarse mutuamente.
* Los métodos del codo y gradiente coincidieron en 3 clusters como el número óptimo.

Cada cluster representa un perfil humano distinto: motivado, intermedio y vulnerable.

**8. Conclusiones y reflexiones**

Este análisis confirma que los datos tienen patrones claros y significativos. Los clusters no son simples divisiones matemáticas, son representaciones de perfiles humanos con diferentes niveles de motivación, empoderamiento y apoyo social. Donde encontramos:

* Un grupo motivado y empoderado puede asumir programas avanzados o retos mayores.
* Un grupo intermedio requiere refuerzo puntual para equilibrar sus áreas débiles.
* Un grupo vulnerable debe ser priorizado con estrategias de acompañamiento.

En resumen, el clustering nos ayuda a ver lo invisible, nos permite descubrir patrones ocultos que guían la toma de decisiones más humanas, efectivas y personalizadas.

**Código:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans, AgglomerativeClustering

from scipy.cluster.hierarchy import dendrogram, linkage

df= d.read\_csv("C:/Users/usuario/Downloads/cervical+cancer+behavior+risk/sobar-72.csv")

print("Dimensiones del dataset:", df.shape)

display(df.head())

X = df.select\_dtypes(include=[np.number])

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

print("\nDescripción estadística básica:")

display(df.describe().T)

pca = PCA(n\_components=2)

X\_pca = pca.fit\_transform(X\_scaled)

print("\nVarianza explicada po

r cada componente principal:")

print(pca.explained\_variance\_ratio\_)

plt.figure(figsize=(6,5))

plt.scatter(X\_pca[:,0], X\_pca[:,1], alpha=0.7)

plt.title("PCA (2 componentes principales)")

plt.xlabel("PC1")

plt.ylabel("PC2")

plt.show()

inertia = []

K = range(1,10)

for k in K:

km = KMeans(n\_clusters=k, random\_state=42)

km.fit(X\_scaled)

inertia.append(km.inertia\_)

plt.figure(figsize=(6,5))

plt.plot(K, inertia, marker='o')

plt.title("Método del Codo - KMeans")

plt.xlabel("Número de clusters")

plt.ylabel("Inercia")

plt.show()

linkage\_matrix = linkage(X\_scaled, method='ward')

plt.figure(figsize=(10,6))

dendrogram(linkage\_matrix, truncate\_mode="lastp", p=12, leaf\_rotation=45, leaf\_font\_size=10, show\_contracted=True)

plt.title("Dendrograma - Clustering Jerárquico")

plt.xlabel("Índice de observaciones")

plt.ylabel("Distancia")

plt.show()

k\_opt = 3

kmeans = KMeans(n\_clusters=k\_opt, random\_state=42).fit(X\_scaled)

df["cluster\_kmeans"] = kmeans.labels\_

agg = AgglomerativeClustering(n\_clusters=k\_opt, linkage="ward").fit(X\_scaled)

df["cluster\_hier"] = agg.labels\_

plt.figure(figsize=(6,5))

plt.scatter(X\_pca[:,0], X\_pca[:,1], c=df["cluster\_kmeans"], cmap="tab10", alpha=0.7)

plt.title("Clusters KMeans (visualización en PCA)")

plt.xlabel("PC1")

plt.ylabel("PC2")

plt.show()

profiles = df.groupby("cluster\_kmeans").mean()

print("\nPerfiles promedio por cluster (KMeans):")

display(profiles)

print("\n Insights obtenidos:")

print("1. El PCA muestra que la variabilidad principal está concentrada en 2 ejes, lo que facilita la visualización.")

print("2. El método del codo sugiere que 3 clusters capturan bien la estructura de los datos.")

print("3. El dendrograma jerárquico también muestra divisiones claras en torno a 3 grupos.")

print("4. Los perfiles promedio indican diferencias claras en variables clave entre los clusters.")

print("5. Los clusters representan distintos perfiles: motivado/empoderado, intermedio y vulnerable.")