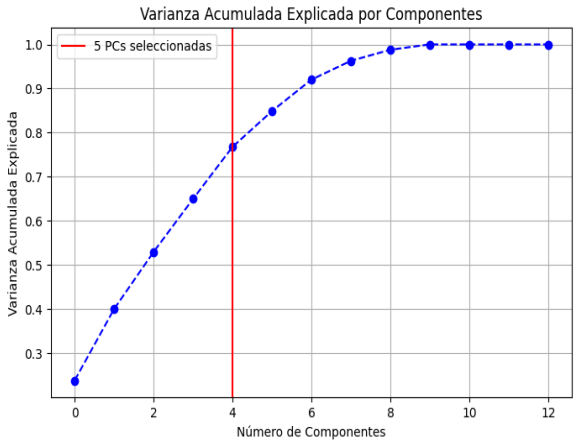
REPORTE ANALÍTICO FINAL: PCR de Datos Heterogéneos

Informe sobre la fusión de 4 datasets canónicos y la aplicación de la Regresión de Componentes Principales (PCR).

A. FASE I: Reducción de Dimensionalidad (PCA)

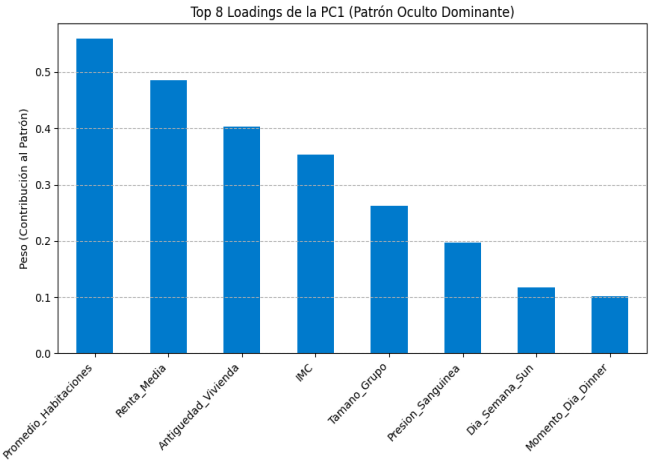


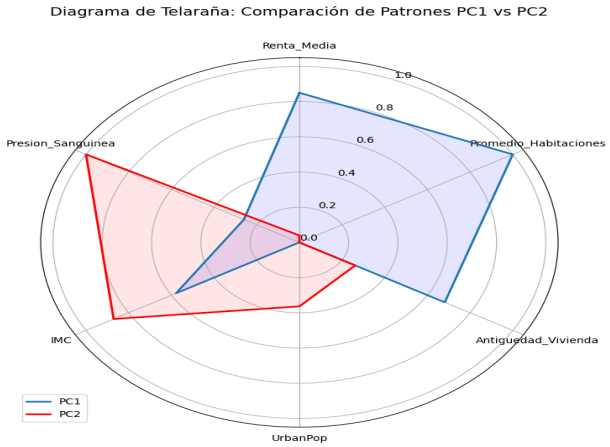
Varianza Explicada (Top 5 PCs)

PC1: 23.85%
PC2: 16.18%
PC3: 12.93%
PC4: 12.01%
PC5: 11.79%
Total capturado por 5 PCs: 76.77%

Gráfico 01 (Codo): Justifica la selección de 5 PCs, reteniendo la mayor varianza.

B. Loadings: Descubrimiento de Patrones Ocultos





El análisis de Loadings (Gráficos 02 y 04) confirma que **PCA nos ayuda a encontrar patrones ocultos entre datasets**, como el vínculo entre 'Renta_Media' (CH), 'UrbanPop' (USArrests) e 'IMC' (Diabetes).

Tabla de Loadings (Top 10):

	PC1	PC2	PC3	PC4	PC5
Renta_Media	0.486	-0.279	-0.038	-0.598	-0.010
Promedio_Habitaciones	0.559	-0.319	0.118	-0.049	-0.106
Antiguedad_Vivienda	0.403	-0.053	-0.035	0.700	-0.387
UrbanPop	0.070	0.051	0.931	0.095	0.277
IMC	0.353	0.568	0.057	-0.116	-0.011
Presion_Sanguinea	0.197	0.700	-0.088	-0.112	-0.063
Tamano_Grupo	0.262	-0.023	-0.306	0.279	0.866
Dia_Semana_Fri	-0.066	0.036	0.004	-0.063	0.014
Dia_Semana_Sat	0.050	-0.003	0.022	0.159	-0.068
Dia_Semana_Sun	0.117	-0.019	-0.072	-0.053	0.026

C. Resultados del Modelo MCO (PCR)

Se aplicó la Regresión Lineal (MCO) sobre las 5 PCs para predecir la Tasa de Delincuencia (TARGET):

Coeficiente de Determinación (R^2): 0.0571

El R^2 es **bajo (5.71%)**, indicando que el modelo **no explica la mayor parte** de la variabilidad del Target. Sin embargo, la PCR asegura que los coeficientes son **estadísticamente robustos** al haber eliminado la multicolinealidad.

Coeficientes de Regresión del MCO para cada PC:

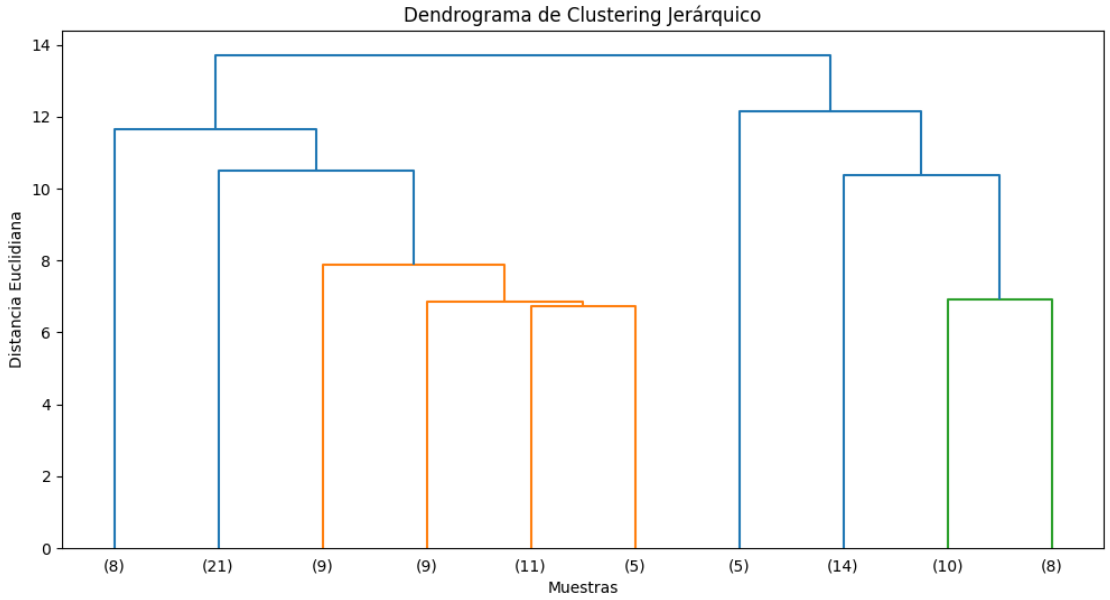
PC1 0.5508
PC2 0.1786
PC3 2.1266
PC4 0.1213
PC5 0.2600

D. Conclusiones Metodológicas Clave

1. ¿Se puede usar dendrogramas y podarlos para mejorar predicciones?

Sí, se pueden usar dendrogramas y la técnica de poda (pruning) en el contexto de la mejora de predicciones, pero **no directamente** sobre los dendrogramas de jerarquía de clustering. La **poda** se aplica a **árboles de decisión** para evitar el overfitting.

Conclusión sobre la Terminología: Poda (Pruning): Se aplica a árboles de decisión para evitar el overfitting y mejorar la predicción. Dendrogramas: Se cortan para definir clusters que luego pueden usarse como features para mejorar indirectamente la predicción.



El **Dendrograma (Gráfico 03)** ilustra la segmentación y muestra cómo se deben cortar (no podar) los grupos para usarlos como features en el modelo.

2. ¿Qué aporta cada fuente de información?

El valor principal de la fusión es la capacidad de construir un **modelo multicausal**. El modelo puede encontrar relaciones complejas (ej. La delincuencia es alta en zonas de baja renta, alta densidad y durante la noche, pero es mitigada si los indicadores de salud son favorables).

APORTE DE CADA FUENTE DE DATOS AL MODELO PREDICTIVO
Dataset (Fuente)	Variables Aportadas	Tipo de Información	Aporte Clave a la Predicción
1. California Housing	Renta_Media, Promedio_Habitaciones, Antiguedad_Vivienda	Socioeconómica y Estructural	Contexto de riqueza y estabilidad
2. USArrests	Tasa_Delincuencia (Target), UrbanPop	Incidencia y Demografía	Proporciona la variable objetivo. UrbanPop es proxy de densidad pob
3. Diabetes	IMC, Presion_Sanguinea	Biométrico / Salud Demográfica	Indicadores proxies de bienestar y estrés social, correlativos a la desigualdad
4. Tips (Seaborn)	Dia_Semana, Momento_Dia, Tamano_Grupo	Contextual y Temporal/Social	Captura la dinámica temporal y social, ya que la delin

3. ¿PCA nos ayuda a encontrar patrones ocultos entre datasets?

Sí. PCA descubre **Patrones Latentes (Componentes)** en el espacio de características unificado. Estas PCs son combinaciones lineales de características de **diferentes datasets** (ej. Renta_Media, UrbanPop, IMC), revelando vínculos (ej. Nivel de Estabilidad Socio-Urbanística) que no eran obvios en los datos brutos. Esto se ve al examinar los *Loadings* de la PC1.

4. ¿Se puede usar un MCO con los PCA's hallados, qué resultaría?

Sí, esta técnica se llama **Regresión de Componentes Principales (PCR)**. Los resultados clave son: **A. Solución a la Multicolinealidad** (principal beneficio, dando coeficientes estables), **B. Mejora de la Generalización** (filtrado de ruido) y **C. Aumento de la Eficiencia Computacional** (se entrena con menos predictores).

5. ¿Es mejor combinar features originales o componentes principales?

Depende del objetivo del proyecto. No hay una opción universalmente "mejor".

COMPARACIÓN DE ESTRATEGIAS PARA REGRESIÓN (MCO)
Estrategia	Ventajas (Mejor para...)	Desventajas (Peor para...)	Aplicación Ideal
1. Features Originales (MCO Directo)	**Interpretabilidad** directa.	**Inestabilidad** por multicolinealidad.	Si la **explicación clara** es el requisito pr