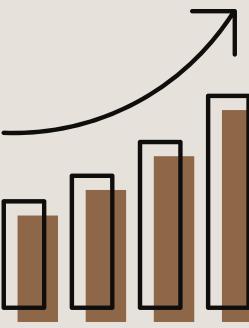


Universidad Externado de
Colombia



PROYECTO: ESTIMACIÓN DE
CONDICIONES DE POBREZA
USANDO VARIABLES
ALTERNATIVAS AL IPM EN
COLOMBIA 2024

Integrantes:
-Yuli Esquivel
-María José Gonzales

Tutor :
Lina Maria Castro

Facultad
Economía



CONTENIDOS

01

Objetivos

02

Descripción

03

Desafíos

04

Entrega de Valor

05

Stakeholders

06

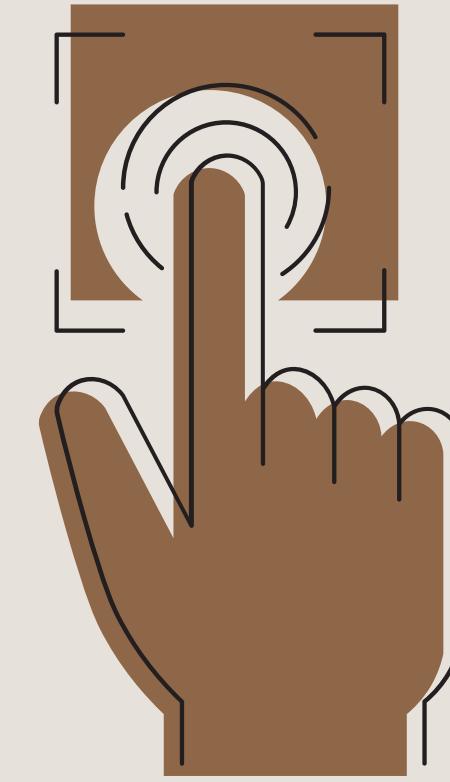
Técnicas que se
utilizarán

07

Variables

08

Fuentes de datos



Organización OBJETIVO

El Índice de Pobreza Multidimensional (IPM) es una medida que captura las privaciones severas que enfrentan las personas simultáneamente en educación, salud y nivel de vida. Sin embargo en Colombia, el DANE calcula este índice, pero su cálculo puede ser complejo.

En este proyecto buscaremos por medio de modelos supervisados predecir la probabilidad de que un hogar/persona sea pobre utilizando variables que no forman parte del IPM tradicional.



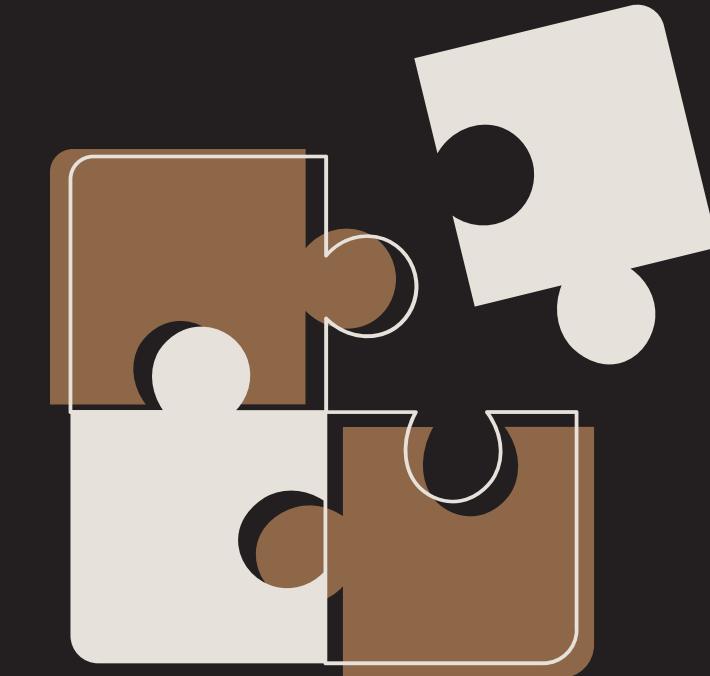
DESCRIPCIÓN

1. Contexto del Proyecto

La pobreza multidimensional es una medida que captura las privaciones de la población en dimensiones como educación, salud, trabajo, juventud y niñez, y condiciones de la vivienda. Predecir y entender este índice a nivel municipal es crucial para la focalización de políticas públicas y la asignación eficiente de recursos.

2. Alcance

El proyecto se centrará en los 32 departamentos de Colombia y el distrito capital . Utilizará los datos del Censo Nacional de Población y Vivienda 2018 y otras fuentes oficiales de años cercanos como línea base. El modelo será transversal (cross-sectional) para el año más reciente con datos completos disponibles.



DESAFÍOS

- Acceso y calidad de los datos (faltantes, inconsistentes, escalas distintas)
- Evitar sesgos que refuercen desigualdades (discriminación por ubicación)
- Selección de variables realmente relevantes (no redundantes con las del IPM)
- Interpretabilidad del modelo (explicar por qué predice lo que predice)

En equipo mejor

ENTREGA DE VALOR

- Las herramientas empleadas serán Python para la construcción de código y pruebas, GitHub como plataforma de control de versiones y colaboración, y Google Colab para la ejecución y documentación interactiva. Se integrarán modelos de IA y bibliotecas estadísticas que permitan análisis reproducibles.
- El proyecto permitirá predecir el Índice de Pobreza Multidimensional (IPM) en departamentos colombianos con precisión, apoyando la toma de decisiones en políticas públicas y focalización de recursos. Se espera fortalecer la capacidad de los entes territoriales para anticipar tendencias, diseñar intervenciones más efectivas y reducir desigualdades.
- Método sostenible: uso de datos abiertos y anónimos, modelos eficientes y reproducibles en Python, control de sesgos con auditorías éticas, versionado en GitHub, monitoreo continuo y documentación clara para transferencia. Garantiza transparencia, impacto social duradero y replicabilidad nacional.



STAKEHOLDERS

Gobierno/entidades públicas

- Ministerio de salud; Ministerio de vivienda, DANE, etc.
- Interés: usar los resultados para políticas públicas, asignación de recursos y programas sociales.

Hogares /ciudadanos

- Los hogares que aportan los datos y que podrían beneficiarse de políticas basadas en los resultados.
- Interés: mejora de servicios, apoyo económico, acceso a programas sociales.

Organizaciones de investigación y académicas

- Universidades o centros de estudios de economía y datos.
- Interés: generar conocimiento, validar métodos, publicar resultados.

Desarrolladores y científicos de datos

- Equipo que construye y prueba los modelos de IA.
- Interés: código reproducible, desempeño del modelo, innovación tecnológica.

TÉCNICAS A USAR

Análisis y exploración de datos

- Limpieza de datos
- Estadística descriptiva
- Visualización de distribuciones
- Gráficos de barras

Modelado y predicción

- Regresión lineal
- Árbol de decisión
- K-Nearset Neighbors



VARIABLES

HOGARES CON ACCESO A INTERNET, POR TIPO DE CONEXIÓN

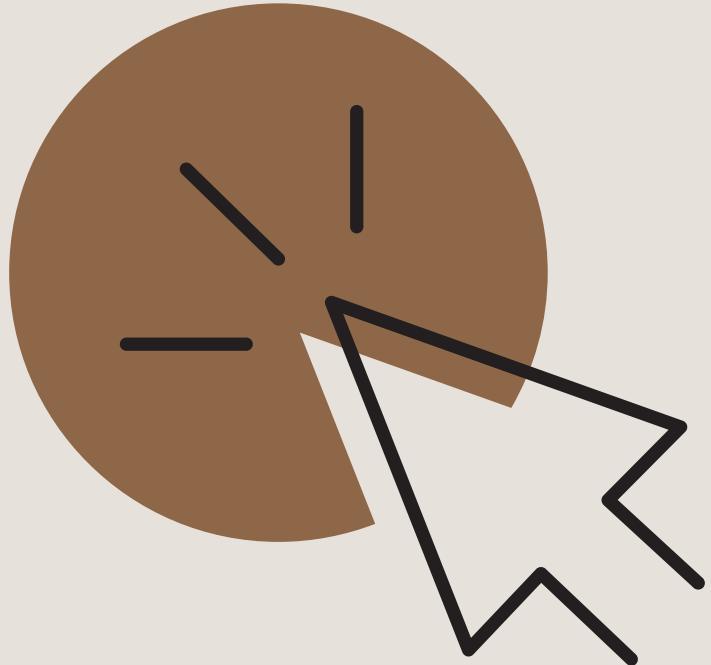
HOGARES POR JEFATURA (MASCULINA O FEMENINA) SIN CÓNYUGE Y CON HIJOS MENORES DE 18 AÑOS

HOGARES POR TIPO DE TENENCIA DE LA VIVIENDA

HOGARES POR ACCESO A SERVICIOS PÚBLICOS, PRIVADOS O COMUNALES

PERCEPCIÓN DE LA CALIDAD DEL SERVICIO DE LA EPS O DE LA ENTIDAD DE SALUD DONDE LAS PERSONAS ESTÁN AFILIADAS

RESULTADOS



- Jefatura del hogar:
- Hombres → sin relación clara
- Mujeres → IPM alto (feminización de la pobreza)
- Servicios públicos:
- Energía, agua, alcantarillado, gas, basuras → mayor cobertura = menor IPM
- Internet:
- Alta cobertura ($>70\%$) → IPM bajo
- Baja cobertura ($<30\%$) → IPM alto
- Vivienda:
- Propia (pagada/en pago) → IPM bajo
- Tenencia precaria (posesión sin título, colectiva) → IPM alto
- Salud (EPS):
- “Buena/Muy buena” → IPM bajo
- “Mala” → IPM alto

FUENTES UTILIZADAS

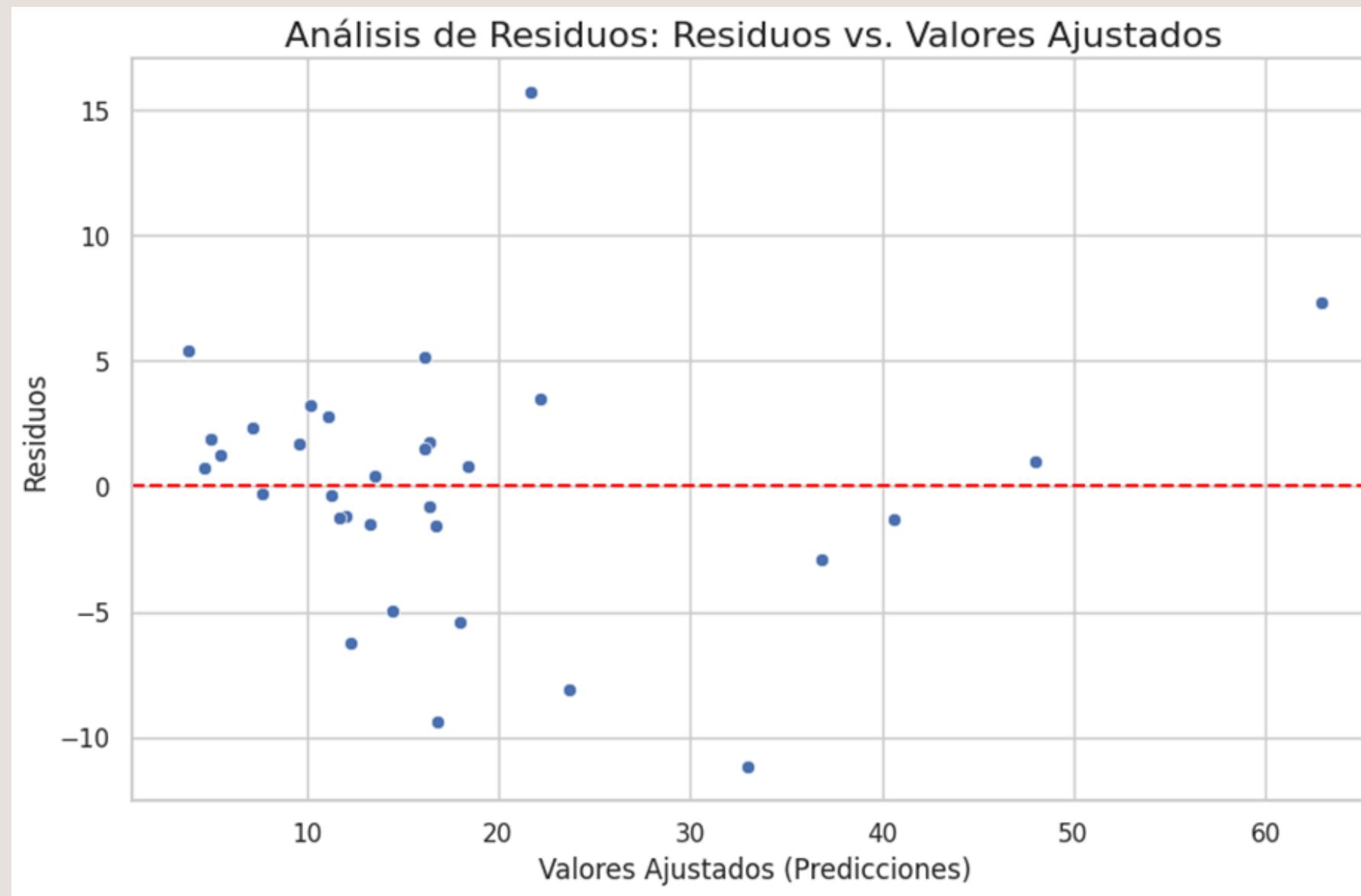
DANE

- **anex-PMultidimensional-Departamental-2024.** Recuperado de
<https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional>
- **anex-ECV-Series-2024.** Recuperado de
<https://www.dane.gov.co/index.php/estadisticas-por-tema/salud/calidad-de-vida-ecv/encuesta-nacional-de-calidad-de-vida-ecv-2024>



REGRESIÓN LINEAL

- Para responder a esta pregunta, empleamos el método estadístico de regresión lineal. Este enfoque nos permitió analizar sistemáticamente la relación entre veintidós variables diferentes y el IPM. La regresión lineal no solo identifica qué variables tienen una conexión significativa con la pobreza, sino que también cuantifica la fuerza de esta relación y establece un nivel de confianza estadística para cada hallazgo.

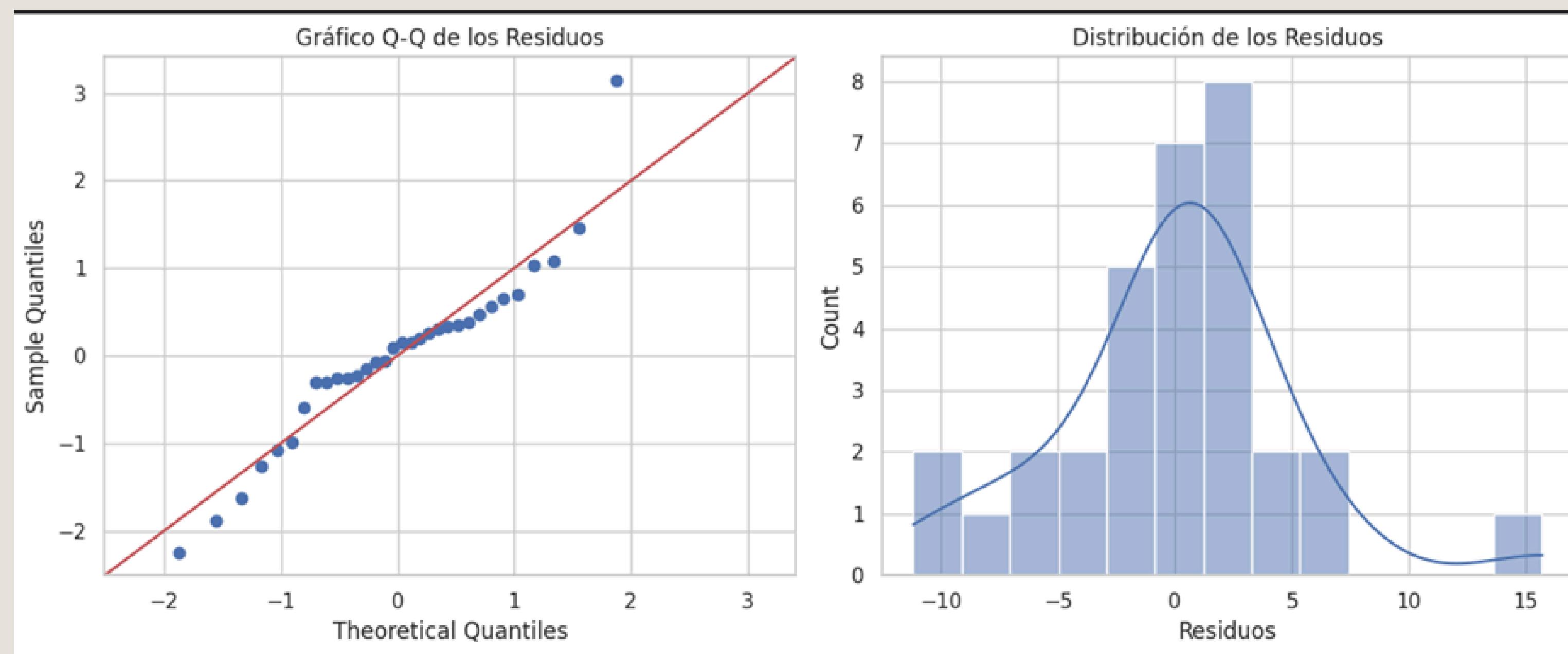


En esta gráfica estamos observando el comportamiento de los errores de nuestro modelo de regresión lineal. Los puntos se distribuyen de forma aleatoria y homogénea alrededor de la línea horizontal.

Lo que esto me indica es que nuestro modelo está funcionando correctamente y de manera consistente. Los residuos que son las diferencias entre los valores reales del IPM y los que predice nuestro modelo- se comportan de forma similar sin importar si estamos prediciendo para municipios con IPM bajo (20 puntos), medio (30-40 puntos) o alto (50-60 puntos).

La dispersión constante que observo a lo largo de todo el eje horizontal significa que la precisión de nuestro indicador de pobreza es comparable en municipios ricos y pobres por igual. No hay zonas donde el modelo cometa errores sistemáticamente más grandes, lo cual es fundamental para la confiabilidad de nuestras estimaciones.

En términos prácticos, esto valida que podemos usar con confianza estadística las variables proxy que identificamos -como acceso a internet y alcantarillado para estimar el nivel de pobreza en cualquier municipio colombiano, sin preocuparnos por que el modelo funcione mejor en unos territorios que en otros.



- Gráfico Q-Q (izquierda)

Los puntos siguen bastante bien la línea diagonal, aunque hay pequeñas desviaciones en los extremos, especialmente en la cola superior.

Los residuos tienen una distribución casi normal, lo que indica que nuestro modelo de regresión lineal es apropiado para los datos. Las ligeras desviaciones en los extremos son normales en datos reales y no afectan la validez del modelo.

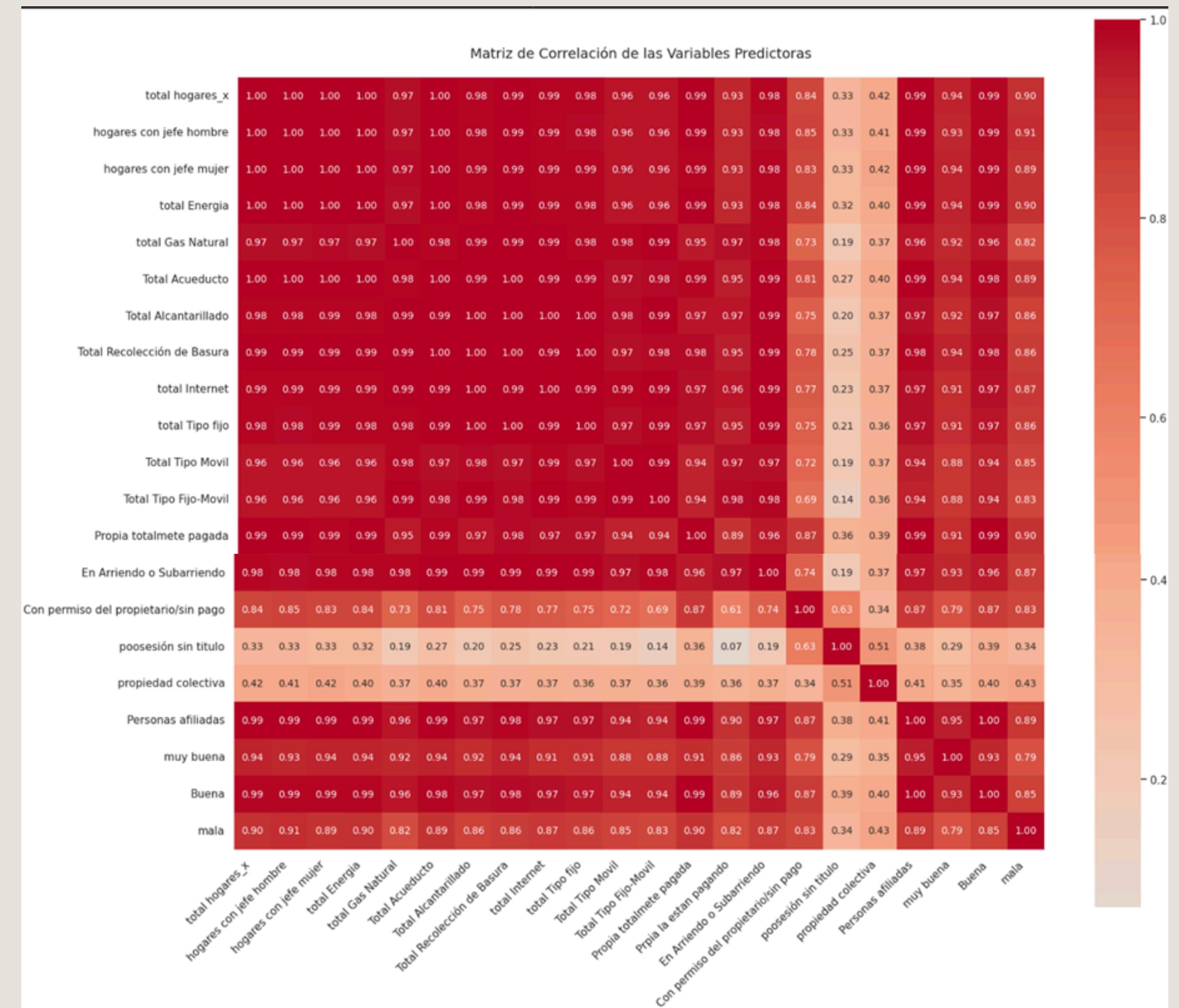
- Histograma de Residuos (derecha)

Lo que observo: La distribución tiene forma aproximadamente acampanada, centrada cerca del cero, con la mayoría de residuos entre -5 y +5. Hay algunos valores más extremos pero con baja frecuencia.

Qué significa: El modelo predice bien en general (residuos cerca de cero) y no tiene sesgos evidentes. La simetría sugiere que el modelo es igualmente preciso para sobreestimar y subestimar el IPM.

En primer lugar, detectamos un problema de multicolinealidad con variables que muestran correlaciones perfectas o casi perfectas (0.97-1.00). Esto es particularmente evidente entre "total hogares", "hogares con jefe hombre" y "hogares con jefe mujer", que básicamente miden lo mismo. Igualmente, los servicios públicos como alcantarillado, recolección de basura e internet presentan correlaciones extremadamente altas, indicando redundancia informativa.

Por otro lado, encontramos variables valiosamente independientes como "posesión sin título" y "propiedad colectiva", que muestran correlaciones bajas (0.19-0.42) con las demás variables. Esta independencia estadística las convierte en candidatas ideales para el modelo, ya que aportan información única no duplicada por otras variables.

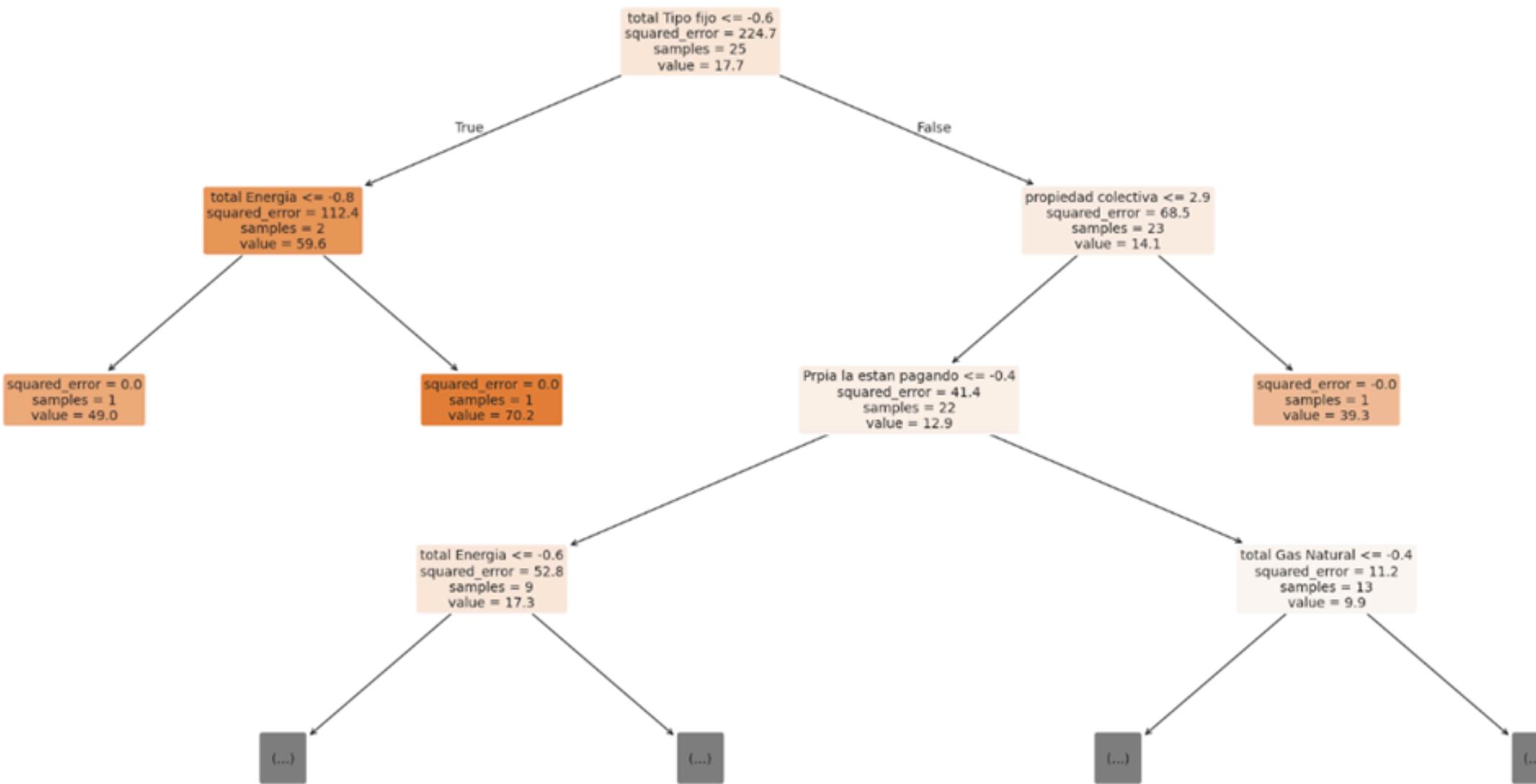


ARBOL DE DECISIÓN

- Se implementó el método de Árboles de Decisión por su capacidad para capturar relaciones no lineales entre variables predictoras y el IPM, identificar puntos de corte óptimos y descubrir interacciones entre variables. Este enfoque genera reglas de decisión interpretables, útiles para la recolección de datos y el diseño de políticas públicas.



Visualización Árbol de Regresión (primeras 3 capas)

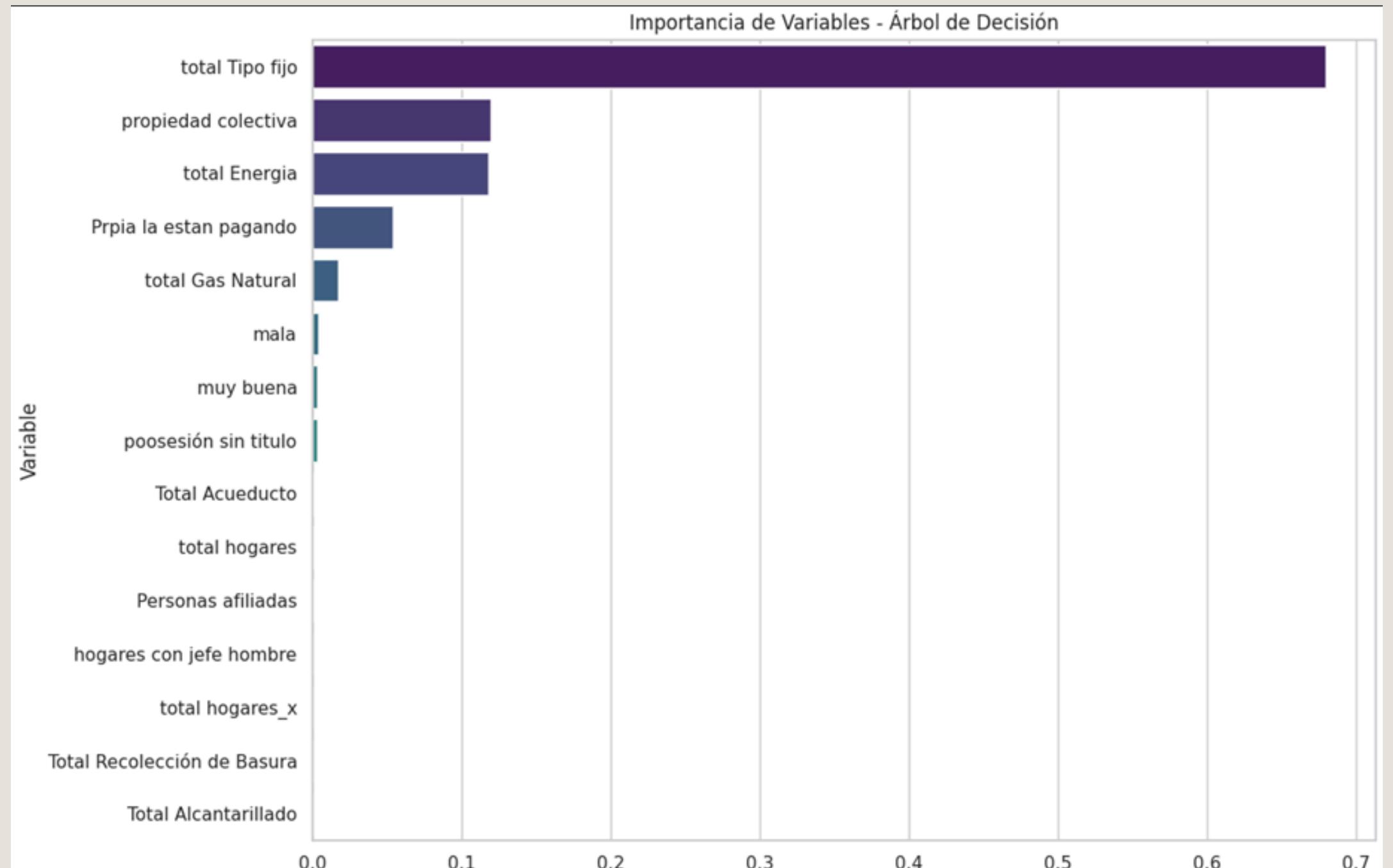


El árbol de decisión comienza con la variable "total Tipo fijo" (servicio telefónico fijo). Si este valor es menor o igual a -0.6, el árbol se divide en dos ramas principales, mostrando que esta variable sola explica gran parte del comportamiento del IPM.

En la primera rama ($\text{Tipo fijo} \leq -0.6$), el árbol utiliza "Total Energía" como siguiente criterio. Cuando el acceso a energía es muy bajo (≤ -0.8), encontramos valores extremos de IPM: en un caso 49.0 y en otro 70.2, que representan niveles de pobreza multidimensional muy altos. Esto revela que la combinación de falta de servicio telefónico y energía eléctrica , identifica territorios con pobreza severa.

En la rama contraria ($\text{Tipo fijo} > -0.6$), el árbol "Propiedad colectiva" y "Propiedad que la estan pagando". Aquí vemos valores de IPM más moderados, entre 12.9 y 14.1, indicando que la seguridad en la tenencia de vivienda se asocia con menores niveles de pobreza.

Los valores de "squared error" nos muestran la precisión del árbol: errores cercanos a cero (0.0) indican vértices donde el modelo predice casi perfectamente, mientras que errores más altos (52.8, 68.5) señalan vértices con mayor variabilidad interna. El árbol demuestra que la pobreza multidimensional sigue patrones predecibles, la falta de servicios básicos telefónicos y energéticos identifica territorios críticos, mientras que las condiciones de vivienda permiten refinar las predicciones en zonas con mejor acceso a servicios.



El análisis revela que tres variables clave explican más del 90% del poder predictivo del modelo:

Total Tipo fijo (0,68%) - Servicio telefónico fijo como mejor predictor individual

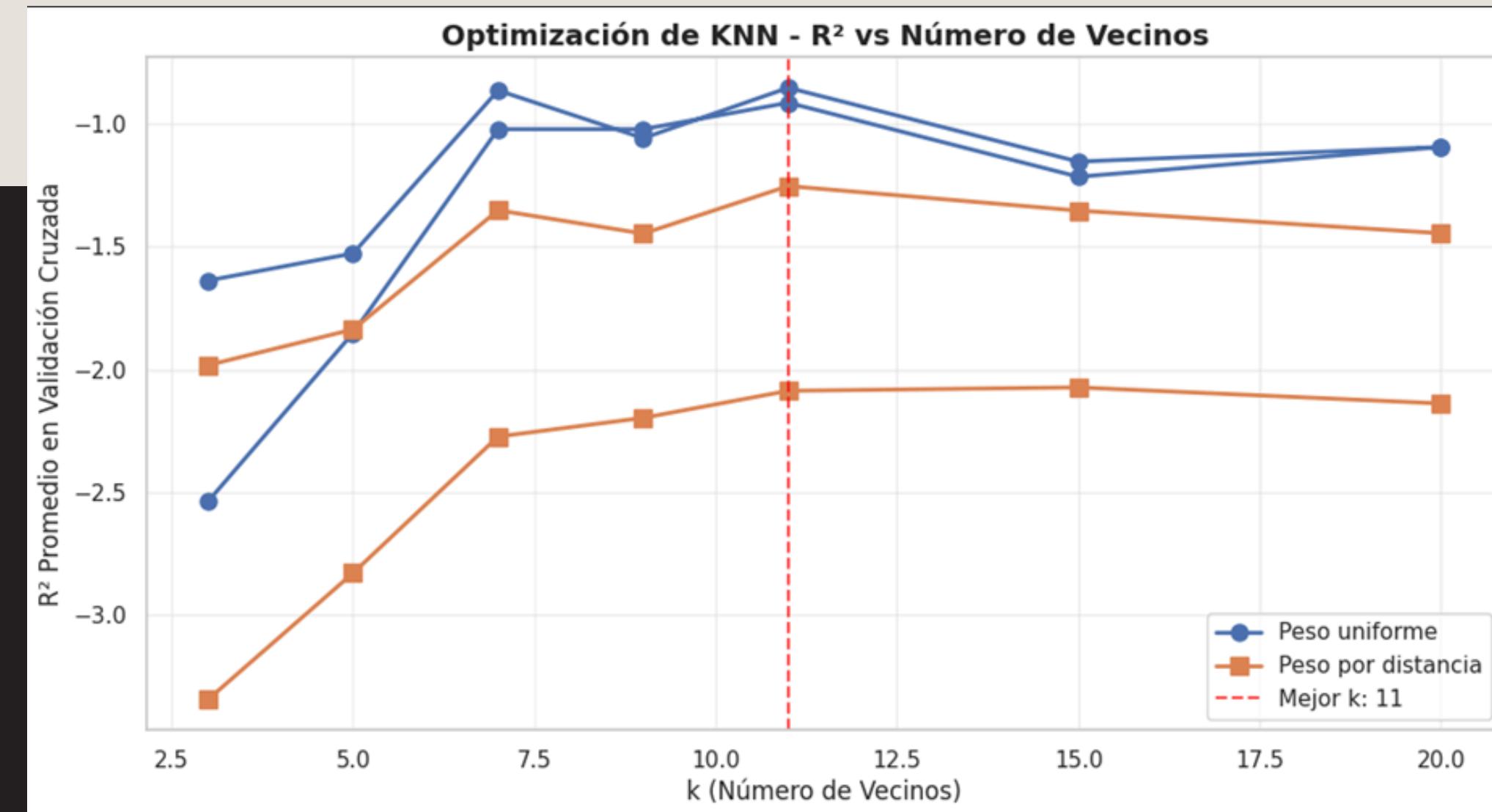
Propiedad colectiva (0,12%) - Tipo de tenencia de vivienda

Total Energía (0,12%) - Acceso a servicio eléctrico
Las demás variables (servicios públicos, salud, demografía) muestran importancia marginal (<5%).

Conclusión práctica: Los municipios pueden optimizar recursos enfocándose principalmente en medir telecomunicaciones fijas, tipo de propiedad y acceso a energía para obtener estimaciones confiables de pobreza con mínimo esfuerzo de recolección.

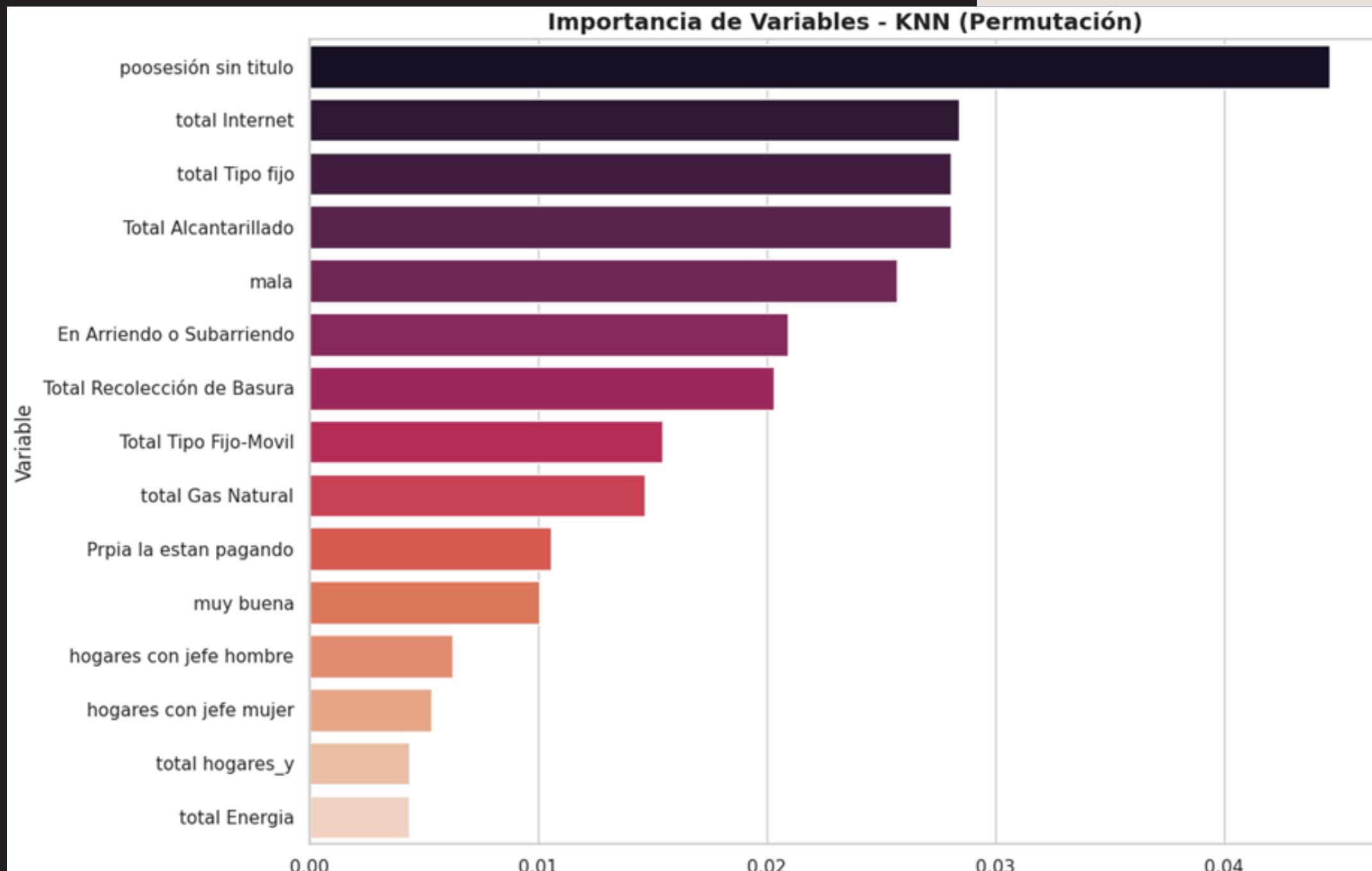
K-NN

- Se implementó el método K-Nearest Neighbors por su capacidad para identificar patrones basados en similitud entre observaciones, aprovechando el principio de que territorios con características socioeconómicas similares deberían presentar niveles de pobreza multidimensional comparables. Esta aproximación permite capturar relaciones complejas sin asumir formas funcionales específicas entre predictores y el IPM, utilizando la proximidad en el espacio multivariado para generar estimaciones



Esta gráfica nos muestra algo preocupante: sin importar cuántos "vecinos" usemos en el modelo KNN, los resultados siempre son peores que hacer una predicción básica.

Los números negativos significan que el modelo no está aprendiendo patrones útiles de los datos. Aunque encontramos que con 11 vecinos se obtiene el "mejor" resultado, sigue siendo malo.



El análisis muestra que "posesión sin título" es la variable más influyente en el modelo KNN, seguida por el acceso a internet y telefonía fijo. Sin embargo, todas las variables presentan una importancia muy baja (valores menores a 0.04), lo que confirma el bajo poder predictivo general de este modelo. Esto refuerza nuestra conclusión anterior: el KNN no es adecuado para estimar la pobreza multidimensional, ya que ninguna variable logra destacarse como un predictor confiable, a diferencia de los otros modelos donde identificamos variables claramente determinantes.

CONCLUSIONES

Este estudio demuestra que sí es viable implementar un sistema de variables proxy para estimar la pobreza multidimensional en Colombia, logrando el objetivo central de agilizar el monitoreo, reducir costos operativos y facilitar análisis más frecuentes, sin comprometer la precisión en la identificación de poblaciones vulnerables. Los resultados identifican el acceso a internet, los servicios de telecomunicaciones y la infraestructura básica como las variables más confiables y prácticas para este fin, ofreciendo a los gobiernos locales una herramienta accesible que optimiza recursos y mejora la focalización de políticas públicas contra la pobreza.

Variables de Alto Impacto:

- Acceso a Internet - Indicador más sólido
- Servicios de Telecomunicaciones - Teléfono fija/móvil
- Infraestructura de Saneamiento - Alcantarillado
- Acceso a Energía - Variable clave
- Condiciones de Vivienda - Tipo fijo y propiedad colectiva

GRACIAS