Determinantes de la Pobreza Monetaria en Colombia: Un Enfoque Integral mediante Ciencia de Datos y Técnicas de Machine Learning

Daniel Gabriel Restrepo Castaño
Código 8992833

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director

Daniel Enrique González Gómez

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 1 DE 2024

FICHA RESUMEN

TÍTULO DEL PROYECTO:

1. ÁREA DE TRABAJO: Análisis socioeconómico y modelado predictivo.
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE: Daniel Gabriel Restrepo Castaño
4. CORREO ELECTRÓNICO:
5. DIRECCIÓN Y TELEFONO:
6. DIRECTOR: Daniel Enrique González Gómez
7. VINCULACIÓN DEL DIRECTOR: Profesor, Departamento de Ciencias Naturales y Matemáticas, Facultad de Ingeniería y Ciencias
8. CORREO ELECTRÓNICO DEL DIRECTOR:
9. CODIRECTOR: NA
10. GRUPO O EMPRESA QUE LO AVALA: NA
11. OTROS GRUPOS O EMPRESAS: NA
12. PALABRAS CLAVE: Pobreza monetaria, ciencia de datos, machine learning, Colombia, predicción de pobreza.
13. FECHA DE INICIO: mayo 2023
14. DURACIÓN ESTIMADA: 12 meses
15. RESUMEN:

Este proyecto aborda la identificación de los principales determinantes de la pobreza monetaria en Colombia mediante un enfoque integral basado en ciencia de datos y técnicas de aprendizaje automático. A partir de un análisis exhaustivo de factores socioeconómicos, demográficos y de vivienda, se desarrollarán modelos predictivos que permitirán identificar hogares en riesgo de caer en la pobreza. Se utilizarán datos de la Gran Encuesta Integrada de Hogares (GEIH) del DANE y otros conjuntos de datos, aplicando técnicas avanzadas de análisis y modelado para mejorar la comprensión de la pobreza. Los resultados proporcionarán evidencia clave para el diseño de políticas públicas más focalizadas y efectivas en la reducción de la pobreza monetaria en el país.

Contenido

INTRODUCCIÓN

En el presente proyecto se abordó el desafío de desarrollar una medida integral que permitiera identificar y caracterizar las condiciones de pobreza y vulnerabilidad de manera combinada a nivel territorial. Si bien existen múltiples indicadores oficiales como el Índice de Pobreza Multidimensional (IPM), el Índice Multidimensional de Condiciones de Vida (IMCV) y la pobreza monetaria, su análisis por separado dificulta una comprensión holística de las privaciones que enfrentan los territorios urbanos.

Dado lo anterior, se construyó un Índice Combinado de Pobreza y Vulnerabilidad (ICPV), integrando componentes del IPM, el IMCV y las tasas de pobreza monetaria (moderada y extrema). El índice fue diseñado para ser estadísticamente robusto, con ponderaciones justificadas tanto desde la igualdad como desde criterios derivados por Análisis de Componentes Principales (PCA) y orientado a reflejar de forma coherente la realidad socioeconómica de los territorios.

Aunque el planteamiento original del proyecto no restringía el análisis a una ciudad específica, se realizó un foco en la ciudad de Medellín debido a la alta calidad de sus datos disponibles por comuna, la trazabilidad histórica, y el acceso directo a fuentes oficiales como el DANE y el Departamento Administrativo de Planeación Distrital. Esto permitió una construcción metodológica precisa y una validación contextual sólida.

Como parte del análisis, se aplicaron técnicas estadísticas de preprocesamiento, selección y ponderación de variables, seguido de la construcción del índice. Posteriormente, se realizó una segmentación de comunas mediante análisis de clústeres y se exploró el uso de modelos de machine learning (Random Forest) para validar la importancia relativa de las variables. Finalmente, se generaron mapas temáticos y comparaciones con otros indicadores del mercado laboral para validar la coherencia territorial del ICPV.

Los resultados muestran que el ICPV permitió identificar con claridad comunas con mayores niveles de pobreza acumulada (como Popular, Santa Cruz y Manrique) y comunas con menores niveles (como El Poblado y Laureles). Además, se evidenció una alta correlación entre el ICPV y los indicadores laborales, como la tasa de informalidad y el desempleo, lo que valida su aplicabilidad como herramienta para el diseño de políticas públicas focalizadas.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Definición del problema

La pobreza monetaria, según la definición del DANE, se refiere a la situación en la que los ingresos per cápita de un hogar son inferiores a la línea de pobreza, establecida con base en el costo de una canasta básica de bienes y servicios esenciales. Más del 29% de la población colombiana aún vivía en condiciones de pobreza monetaria, según cifras del DANE (2022).

Este concepto abarca no solo los bajos ingresos, sino también limitaciones en el acceso a la educación, la salud, el empleo formal y la vivienda digna. A pesar de los avances en la reducción de la pobreza, persistían desigualdades significativas que dificultaban el progreso hacia un desarrollo inclusivo. Este proyecto abordó estas limitaciones a través de un enfoque integral que aplicó ciencia de datos y técnicas de aprendizaje automático para identificar y caracterizar los factores que inciden en la pobreza.

1.2. Planteamiento del problema

La pobreza es uno de los desafíos más persistentes y complejos que enfrentaban Medellín y Colombia en general [1]. A pesar de los esfuerzos significativos para reducirla, un porcentaje considerable de la población colombiana (29.6% en 2022) seguía viviendo en condiciones de pobreza monetaria [1]. La medición tradicional de la pobreza monetaria en Colombia se basaba en la comparación del ingreso per cápita de los hogares con una línea de pobreza establecida [2]. Sin embargo, este enfoque no capturaba completamente la complejidad de los factores que contribuían a la pobreza [3].

La pobreza monetaria no era solo una cuestión de bajos ingresos; también estaba influenciada por una variedad de factores socioeconómicos, demográficos y de vivienda. Estos factores incluían, pero no se limitaban al acceso a la educación y la salud, las condiciones de vivienda, la situación laboral y las características demográficas de los hogares. La ciencia de datos permitió identificar y cuantificar la influencia de estos determinantes a través del análisis de grandes conjuntos de datos, como encuestas de hogares, registros administrativos y datos geoespaciales.

Por ejemplo, se utilizaron modelos de aprendizaje automático para analizar patrones de vulnerabilidad en función de diversas variables, lo cual permitió identificar a las comunas más expuestas a múltiples formas de pobreza y diseñar indicadores más integrales. Además, el análisis de datos reveló patrones y relaciones ocultas entre diferentes factores, lo que condujo a nuevas hipótesis y enfoques para abordar la pobreza.

Una comprensión más profunda de estos determinantes, facilitada por el uso de herramientas y técnicas de la ciencia de datos, proporcionó una base sólida para diseñar políticas públicas más efectivas y focalizadas que abordaran no solo los síntomas, sino también las causas subyacentes de la pobreza.

1.3. Formulación del problema

La pregunta central que guía este proyecto es:

¿Cuáles son los principales determinantes de la pobreza monetaria en Colombia y cómo pueden las técnicas avanzadas de ciencia de datos, en particular el machine learning, mejorar la comprensión de la interacción entre estos factores y predecir con mayor precisión la probabilidad de que un hogar caiga en condiciones de pobreza?

1.3.1. Sistematización

Para responder de manera clara y precisa a la pregunta principal del proyecto, se han formulado las siguientes subpreguntas, que permiten profundizar en los distintos aspectos que afectan a la pobreza monetaria.

¿Cuáles son los principales factores socioeconómicos (ingresos, empleo, educación), demográficos (tamaño del hogar, edad, género) y de vivienda (calidad, acceso a servicios básicos) que impactan significativamente la pobreza monetaria en Colombia y cómo varía su influencia entre diferentes grupos poblacionales?

¿Cómo pueden las técnicas de aprendizaje automático, incluyendo modelos de clasificación y regresión ser utilizadas para identificar patrones ocultos entre los determinantes de la pobreza y mejorar la predicción de la probabilidad de que un hogar caiga en la pobreza?

¿De qué manera la información contenida en fuentes de datos no convencionales, como redes sociales, datos de movilidad o imágenes satelitales, puede contribuir a entender el concepto de pobreza monetaria y mejorar la capacidad predictiva de los modelos de machine learning?

¿Cómo se pueden traducir los resultados de los modelos predictivos en recomendaciones concretas para el diseño de políticas públicas que abordan de manera más efectiva las causas fundamentales de la pobreza en Colombia?

Estas subpreguntas permitirán descomponer el problema principal en componentes más específicos, facilitando el análisis detallado de los factores que inciden en la pobreza y orientando el desarrollo de las políticas públicas.

1.4. Objetivos

1.4.1. Objetivo General

Determinar los principales determinantes de la pobreza monetaria utilizando un enfoque integral que considere factores socioeconómicos, demográficos y de vivienda, y desarrollar modelos predictivos mediante técnicas de machine learning para identificar de manera precisa los hogares en riesgo de caer en pobreza.

1.4.2. Objetivos Específicos

Caracterizar las condiciones de vivienda y las características demográficas de los hogares pobres y no pobres para identificar su impacto en la pobreza monetaria, utilizando análisis de clustering para discernir patrones.

Analizar la relación entre las características del hogar (tamaño, acceso a servicios básicos, hacinamiento), las características demográficas (edad, sexo, estado civil), el acceso a la educación y la salud, y la situación laboral (empleo, desempleo, tipo de empleo) con la pobreza monetaria, utilizando modelos de regresión y otras técnicas apropiadas.

Desarrollar un modelo predictivo integral que utilice técnicas de machine learning, destacando la aplicación de algoritmos como Random Forest, para identificar hogares en riesgo de caer en pobreza.

Implementar un proceso de validación para evaluar la precisión y la aplicabilidad del modelo predictivo en la identificación de hogares en riesgo de caer en pobreza.

1.5. Marco de Referencia

1.5.1. Marco Teórico

El análisis de la pobreza monetaria se sustenta en diversas teorías que abordan los factores que inciden en la pobreza desde diferentes perspectivas. Algunas de las teorías más relevantes para este estudio incluyen:

Tabla 1. Teorías con incidencia en la pobreza y vulnerabilidad

Teoría/Enfoque	Descripción	Referencia
Teoría del Capital Humano	Relación la pobreza con la falta de inversión en educación y salud, disminuyendo las oportunidades laborales.	[4]

Teoría de la Segmentación del Mercado Laboral	El mercado laboral está segmentado en niveles con diferente remuneración y estabilidad, dificultando la movilidad.	[5]
Teoría de la Nueva Geografía Económica	Enfatiza cómo la ubicación geográfica influye en la pobreza; las regiones con menor acceso a infraestructura tienen mayores tasas de pobreza.	[6]
Enfoque de las Capacidades (Amartya Sen)	Defina la pobreza como una privación de capacidades básicas para llevar una vida plena, más allá de la falta de ingresos.	[7]
Enfoque Multidimensional de la Pobreza	Reconocer que la pobreza es multifacética, considerando factores como educación, salud, vivienda y servicios básicos.	[8]

Fuente: Elaboración propia.

Este marco conceptual fundamenta la construcción del Índice Combinado de Pobreza y Vulnerabilidad (ICPV), que integra datos de pobreza monetaria, IPM y condiciones de vida. Las técnicas estadísticas y de aprendizaje automático empleadas permitieron identificar patrones complejos y generar un indicador robusto para la toma de decisiones a nivel territorial.

1.5.2. Antecedentes

Los estudios sobre la pobreza en Colombia han identificado varios determinantes claves que inciden significativamente en esta problemática. Uno de los factores más relevantes es el impacto de la educación. La falta de acceso a una educación de calidad limita las oportunidades laborales y perpetúa la pobreza intergeneracional. Investigaciones como las de Cárdenas y Bernal [9] señalan que mejorar la cobertura y calidad educativa es fundamental para reducir los niveles de pobreza, ya que influye directamente en la empleabilidad y los ingresos futuros de las personas.

Otro factor determinante es el empleo precario y el desempleo. La informalidad laboral, caracterizada por trabajos sin estabilidad ni beneficios, incrementa el riesgo de caer en condiciones de pobreza. Bermúdez y Pérez [10] destacan que esta situación afecta especialmente a las mujeres y a los jóvenes, quienes enfrentan barreras significativas para acceder a empleos formales y bien remunerados. Esto genera un ciclo de vulnerabilidad que se perpetúa en los hogares.

Las condiciones de vivienda también juegan un papel crucial. La precariedad habitacional, manifestada en el hacinamiento, la falta de servicios básicos y el uso de materiales inadecuados en la construcción, tiene un impacto directo en el bienestar de las familias. Según Franco y López-

Calva [11], la mejora en las condiciones de vivienda no solo contribuye al confort y la seguridad de los hogares, sino que también influye en otros indicadores de calidad de vida, como la salud y el rendimiento educativo. En comunas como Villa Hermosa, San Javier y Robledo, los indicadores del Índice Multidimensional de Condiciones de Vida (IMCV) reflejan rezagos importantes en calidad de infraestructura, acceso a servicios públicos y entorno residencial, lo que se asocia con mayores niveles de pobreza multidimensional.

Además, el DANE ha reportado que para el año 2024 la incidencia de pobreza multidimensional fue del 11,5%, lo cual evidencia que, si bien ha habido avances, persisten brechas importantes entre territorios [12]. En esa misma línea, el Banco Mundial ha señalado que más de 16 millones de personas en Colombia viven en situación de pobreza, con mayores afectaciones en departamentos históricamente rezagados. Entre los factores más determinantes se encuentran la informalidad laboral y la pobreza de aprendizaje [13]. A nivel regional, la CEPAL ha advertido que, a pesar de la recuperación post-pandemia, las tasas de pobreza extrema siguen siendo altas, especialmente entre mujeres y jóvenes [14].

Este proyecto buscó ampliar estos antecedentes mediante la aplicación de técnicas de ciencia de datos, análisis factorial y aprendizaje automático. A través de la integración del IPM, el IMCV y las tasas de pobreza monetaria, se desarrolló un Índice Combinado de Pobreza y Vulnerabilidad (ICPV) para las comunas de Medellín, el cual permitió una clasificación territorial más robusta y coherente con los datos observados. Además, mediante técnicas como el Análisis de Componentes Principales (PCA) y el algoritmo de agrupamiento K-means, se identificaron clústeres de comunas con patrones similares de pobreza, lo cual facilitó la comparación entre territorios y la identificación de comunas críticas para la acción institucional.

Estos hallazgos aportan a una mejor comprensión de la pobreza en contextos urbanos complejos, y permiten orientar estrategias de intervención diferenciadas. En particular, el enfoque multivariable y la validación del ICPV contra indicadores laborales reales refuerzan su utilidad como herramienta para la planeación social, el monitoreo de brechas y la formulación de políticas públicas más focalizadas y basadas en evidencia.

2. METODOLOGÍA

Este capítulo describe detalladamente el enfoque metodológico seguido para la construcción del Índice Combinado de Pobreza y Vulnerabilidad (ICPV), así como los procedimientos estadísticos, analíticos y computacionales utilizados para responder a los objetivos propuestos en el presente proyecto. La metodología adoptada combina técnicas tradicionales de análisis social con herramientas modernas de la ciencia de datos y del aprendizaje automático, permitiendo una comprensión multidimensional y territorialmente diferenciada de las condiciones de pobreza en las comunas urbanas y territorios rurales de Medellín.

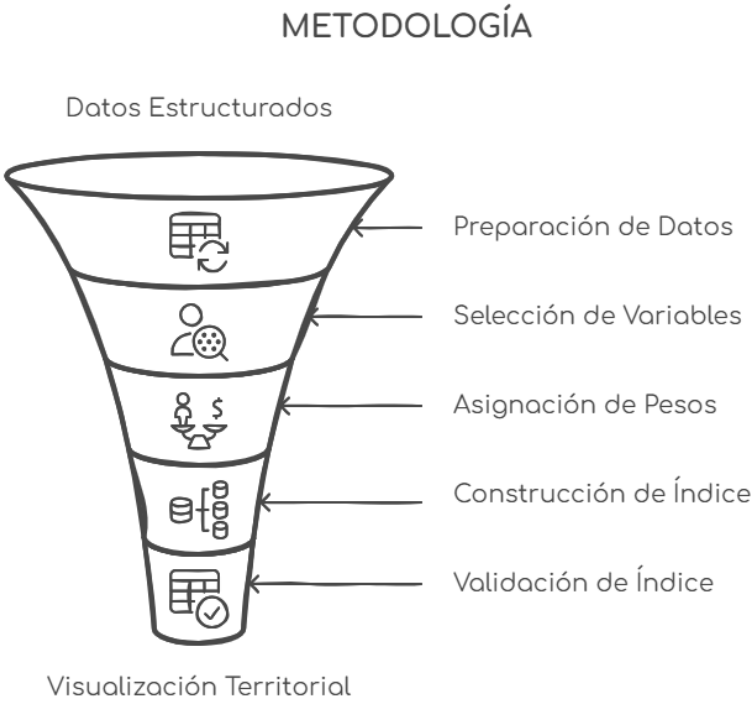

Imagen 1 Esquema metodología del proyecto
Fuente: Elaboración propia.

2.1 Enfoque general del estudio

El estudio se desarrolló bajo un enfoque cuantitativo, de carácter exploratorio y explicativo. A partir de una batería de indicadores socioeconómicos, se construyó un índice sintético capaz de capturar la pobreza y la vulnerabilidad desde una perspectiva multidimensional y territorial. Posteriormente, se aplicaron técnicas estadísticas y de machine learning para validar, clasificar y visualizar los resultados.

El proceso metodológico se basó en cinco etapas centrales: (i) preparación y normalización de datos, (ii) análisis de correlación y reducción de dimensionalidad, (iii) construcción del índice con diferentes esquemas de ponderación, (iv) validación mediante modelos predictivos y clustering, y (v) representación espacial y análisis comparativo de resultados.

2.2 Recolección y estructura de datos

Los datos fueron recolectados a partir de fuentes institucionales disponibles para Medellín, especialmente del Departamento Administrativo de Planeación, con alta desagregación territorial. La información fue consolidada en un único archivo Excel, con las siguientes dimensiones principales:

- Índice de Pobreza Multidimensional (IPM) y sus componentes (2023).
- Índice Multidimensional de Condiciones de Vida (IMCV) y sus dimensiones (2023).
- Incidencia de pobreza monetaria moderada y extrema (2023).
- Variables complementarias como informalidad laboral, trabajo infantil, aseguramiento en salud, desempleo de larga duración, entre otras.
- Identificador de comuna y nombre geográfico correspondiente.

2.3 Preprocesamiento y estandarización de los datos

Durante esta etapa se realizaron procesos de limpieza, transformación y estandarización:

- Homologación de nombres: se corrigieron nombres de columnas con make.names() y se revisó la consistencia semántica.
- Revisión de valores faltantes: se aplicó colSums(is.na()) y summary() para detectar y manejar posibles vacíos o inconsistencias.
- Normalización min-max: se aplicó una transformación min-max a todas las variables, previa inversión de aquellas donde un valor bajo representaba mayor bienestar (como el IMCV o la percepción de calidad de vida), asegurando que, en todos los casos, un valor más alto reflejara mayor privación.
- Conversión de tipos y estructuras: las variables fueron convertidas a formato numérico y las bases estructuradas para permitir su uso en modelos multivariados y gráficos.

2.4 Selección y justificación de variables

La construcción del ICPV se basó en una selección de nueve variables estratégicas, agrupadas en dimensiones clave de la pobreza: educación, empleo, salud, condiciones de vida y pobreza monetaria. Las variables fueron:

- Bajo logro educativo
- Empleo informal
- Sin aseguramiento a salud
- Trabajo infantil
- Desempleo de larga duración
- Pobreza monetaria moderada

- Pobreza monetaria extrema
- Índice de Pobreza Multidimensional (IPM)
- Índice Multidimensional de Condiciones de Vida (IMCV)

La selección se fundamentó en criterios de relevancia teórica, varianza observada, disponibilidad territorial y correlación estadística. Además, se aplicó un Análisis de Componentes Principales (PCA) como técnica de reducción y validación estructural.

2.5 Asignación de ponderaciones

Se exploraron cuatro esquemas de ponderación para construir el índice:

1. Pesos iguales: cada variable recibió un peso uniforme de 1/9.

2. Pesos derivados de PCA: se utilizó la contribución de cada variable al primer componente principal como base de ponderación.

3. Pesos mixtos: se asignaron pesos diferenciados combinando criterio técnico y conocimiento experto, priorizando variables estructurales.

4. Pesos empíricos vía Random Forest: se calculó la importancia relativa de cada variable mediante un modelo predictivo y se normalizaron los resultados como pesos.

Cada esquema dio lugar a una versión distinta del ICPV, cuya comparación posterior permitió evaluar la robustez y sensibilidad del índice.

2.6 Cálculo del ICPV

El Índice Compuesto de Pobreza y Vulnerabilidad (ICPV) fue construido como una suma ponderada de las variables seleccionadas, siguiendo la fórmula general:

$$ICPV = \sum_{i=1}^{9} x_i \cdot w_i$$ Ecuación (1)

donde xi representa la variable i normalizada, y wi el peso asignado según el esquema seleccionado.

El índice se escaló posteriormente al rango [0, 1] mediante normalización min-max, donde 0 representa menor nivel de pobreza y vulnerabilidad relativa, y 1 representa el mayor nivel.

2.7 Validación estadística mediante aprendizaje automático

Se utilizó el algoritmo Random Forest como herramienta de validación empírica y de análisis de sensibilidad del índice. A partir del modelo entrenado con la versión del índice construida con pesos iguales como variable objetivo, se extrajo la importancia relativa de cada predictor (%IncMSE), identificando qué variables explicaban con mayor fuerza la variación del índice.

Estas importancias se usaron también para construir una versión adicional del índice (ICPV_RF), basada en aprendizaje de máquina.

2.8 Clasificación y categorización del índice

Para facilitar la interpretación del ICPV, se procedió a clasificar su versión normalizada en quintiles, de modo que:

- Quintil 1: comunas con menor nivel de pobreza.
- Quintil 5: comunas con mayor nivel de pobreza y vulnerabilidad.

Esta categorización fue aplicada a todas las versiones del índice (pesos iguales, PCA, mixto y RF) para evaluar la consistencia entre ellas.

2.9 Análisis de clústeres

Con el objetivo de segmentar los territorios según su perfil multidimensional, se aplicó el algoritmo k-means clustering sobre las variables ICPV, IPM, IMCV y pobreza monetaria. Previamente, se utilizó el método del codo (elbow method) para determinar el número óptimo de clústeres, identificando tres grupos significativos.

A cada comuna se le asignó un clúster, y se analizaron los perfiles promedio de cada grupo, permitiendo identificar patrones territoriales diferenciados.

2.10 Visualización geoespacial y comparación metodológica

Finalmente, se unió la base de datos con archivos shapefile de las comunas urbanas de Medellín para representar los resultados del ICPV mediante mapas temáticos. Se empleó ggplot2 con geom_sf para visualizar:

- Las distintas versiones del ICPV.
- La distribución por quintiles.

- Los clústeres territoriales.

También se construyó un panel comparativo con los mapas de las tres versiones principales del índice, facilitando su comparación visual y geográfica. Caracterización de las condiciones de vivienda y características demográficas de los hogares pobres y no pobres.

En este capítulo se describirán las actividades realizadas para identificar las principales características de los hogares en situación de pobreza y aquellos que no lo están. El análisis de datos incluye la utilización de técnicas de clustering para descubrir patrones que puedan ayudar a diferenciar estos hogares en función de variables como el acceso a servicios básicos, el tamaño de la vivienda, y las características demográficas (edad, género, estructura familiar, entre otros). Los resultados de este análisis proporcionarán una comprensión más profunda de las condiciones que influyen en la pobreza monetaria en Colombia.

3. RESULTADOS EN FUNCIÓN DEL OBJETIVO GENERAL

El objetivo general de este proyecto fue determinar los principales determinantes de la pobreza monetaria en Colombia mediante un enfoque integral que considera factores socioeconómicos, demográficos y de vivienda, y desarrollar modelos predictivos con técnicas de machine learning para identificar con precisión los hogares en riesgo de caer en pobreza.

Para ello, se aplicó la metodología propuesta al contexto de Medellín, dada su complejidad territorial y la disponibilidad de datos robustos y desagregados a nivel de comuna y corregimientos. Esta elección permitió operacionalizar variables claves relacionadas con educación, empleo informal, acceso a servicios de salud, pobreza monetaria y otros indicadores compuestos como el Índice Multidimensional de Condiciones de Vida (IMCV) y el Índice de Pobreza Multidimensional (IPM).

A partir del conjunto de datos seleccionados y preprocesados, se construyó un índice sintético denominado Índice Combinado de Pobreza y Vulnerabilidad (ICPV), que sintetiza las dimensiones estructurales de la pobreza desde múltiples perspectivas. Para validar y potenciar la capacidad predictiva de este índice, se implementaron técnicas avanzadas de aprendizaje automático, específicamente el algoritmo Random Forest.

El modelo de Random Forest permitió identificar y jerarquizar la importancia relativa de cada variable en la predicción de la pobreza, confirmando que variables como el IPM, el IMCV, la informalidad laboral y la pobreza monetaria moderada tienen un peso significativo en la determinación del riesgo de pobreza. Además, la alta correlación entre el índice ICPV y la versión predictiva derivada del modelo (ICPV_RF) evidenció la robustez y consistencia del enfoque.

Estos resultados confirman que la integración de técnicas estadísticas y de machine learning no solo mejora la comprensión de los determinantes de la pobreza monetaria, sino que también ofrece una herramienta precisa para identificar hogares vulnerables y focalizar políticas públicas con mayor efectividad.

En conclusión, el desarrollo del ICPV y su validación mediante modelos predictivos constituyen un aporte innovador para el análisis territorial de la pobreza en Medellín, con potencial para ser replicado en otros contextos urbanos del país.

4. RESULTADOS EN FUNCIÓN DE LOS OBJETIVOS ESPECÍFICOS

4.1 Caracterización de condiciones demográficas y de vivienda mediante clustering

Objetivo específico: Caracterizar las condiciones de vivienda y las características demográficas de los hogares pobres y no pobres para identificar su impacto en la pobreza monetaria, utilizando análisis de clustering para discernir patrones.

Se aplicó el algoritmo de k-means clustering sobre las comunas de Medellín, utilizando como base el índice ICPV y variables complementarias como el IPM, el IMCV y las tasas de pobreza monetaria. El método del codo indicó que tres clústeres eran óptimos para describir los perfiles territoriales.

Los resultados muestran tres conglomerados claramente diferenciados:

- Clúster 1: Comunas con menor nivel de pobreza, mejores condiciones estructurales y bajos niveles de IMCV e IPM.
- Clúster 2: Comunas con los niveles más altos de pobreza estructural y monetaria.
- Clúster 3: Comunas intermedias, con condiciones mixtas.

Esta segmentación permitió evidenciar la diversidad interna de Medellín en términos de pobreza y vulnerabilidad, y constituye un insumo clave para el diseño de intervenciones focalizadas.

Definición Cualitativa de los Clústeres Territoriales

El análisis de clústeres permitió segmentar las comunas de Medellín en tres grupos con perfiles socioeconómicos claramente diferenciados, facilitando una comprensión más profunda de la distribución territorial de la pobreza y la vulnerabilidad. A continuación, se detalla la definición cualitativa de cada clúster:

Clúster 1: Comunas de Menor Pobreza y Mayor Bienestar. Este conglomerado agrupa a las comunas con las mejores condiciones socioeconómicas. Se caracterizan por tener los promedios más bajos en todos los indicadores de privación: presentan un bajo Índice Combinado de Pobreza y Vulnerabilidad (ICPV) (0.560), una incidencia mínima de Pobreza Multidimensional (IPM) (3.51) y las tasas más bajas de pobreza monetaria. Sus valores en el Índice Multidimensional de Condiciones de Vida (IMCV) son también los más bajos (0.30), lo que refleja mejores condiciones estructurales de vivienda y acceso a servicios. Territorialmente, este clúster corresponde a zonas como El Poblado y Laureles, que consistentemente muestran los menores niveles de carencias.

Clúster 2: Comunas de Alta Pobreza y Vulnerabilidad Crítica. Este grupo identifica a los territorios con la mayor concentración de pobreza y vulnerabilidad estructural y monetaria en la ciudad. Presenta los valores promedio más altos en todos los indicadores analizados: ICPV (2.737), IPM (19.15), pobreza monetaria moderada (0.806) y pobreza monetaria extrema (0.771). Su alto valor de IMCV (0.88) indica una acumulación de carencias estructurales significativas. Este perfil corresponde a las comunas que requieren la mayor atención institucional, como Popular, Manrique y Santa Cruz, donde se evidencian las mayores desigualdades.

Clúster 3: Comunas en Transición con Condiciones Mixtas. Este clúster representa un grupo intermedio de comunas con condiciones socioeconómicas mixtas. Sus indicadores de pobreza y vulnerabilidad se sitúan entre los valores de los clústeres 1 y 2, con un ICPV de 1.607 y un IPM de 10.82. Si bien no enfrentan el nivel de privación crítico del Clúster 2, muestran rezagos y vulnerabilidades importantes que las distinguen de las comunas con mejores condiciones del Clúster 1, posicionándolas como territorios en una situación de transición.

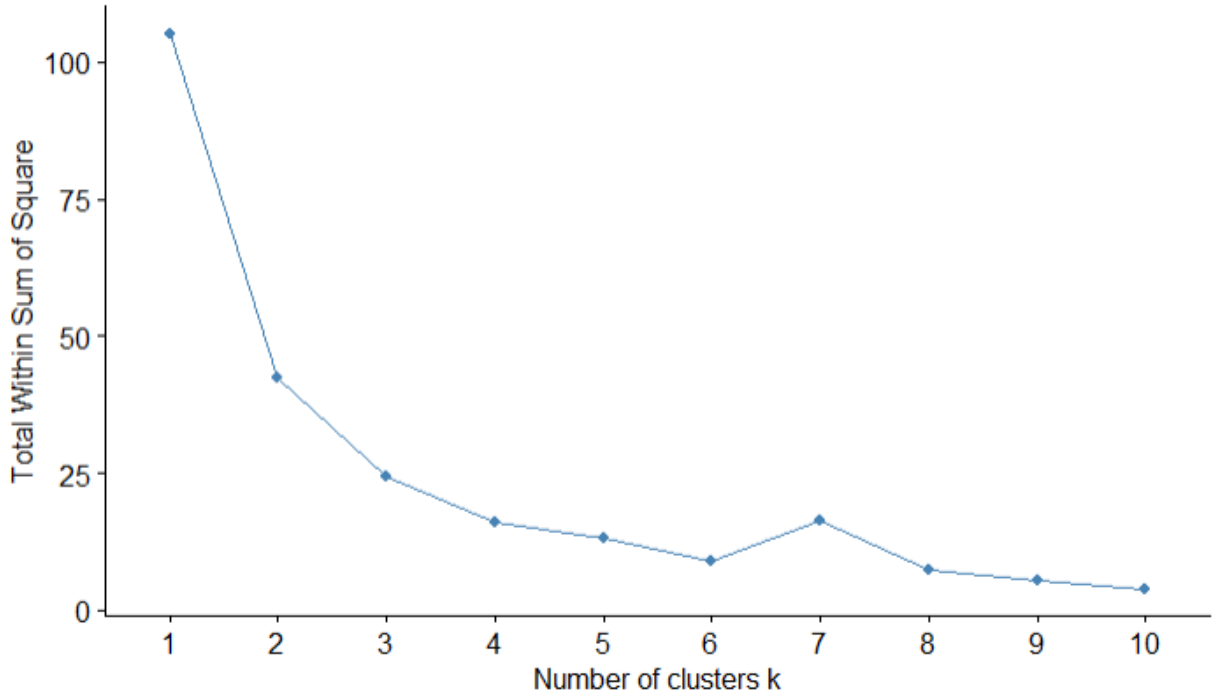

Imagen 2 Método del codo para determinar el número óptimo de clústeres
Fuente: Elaboración propia.

Se utilizó el método del codo (Elbow Method) para determinar el número óptimo de clústeres mediante el análisis de la suma de cuadrados intra-cluster (WSS). El punto de inflexión de la curva, observado en k = 3, sugiere que tres clústeres representan una estructura adecuada para segmentar las comunas de Medellín con base en sus características multidimensionales de pobreza y vulnerabilidad.

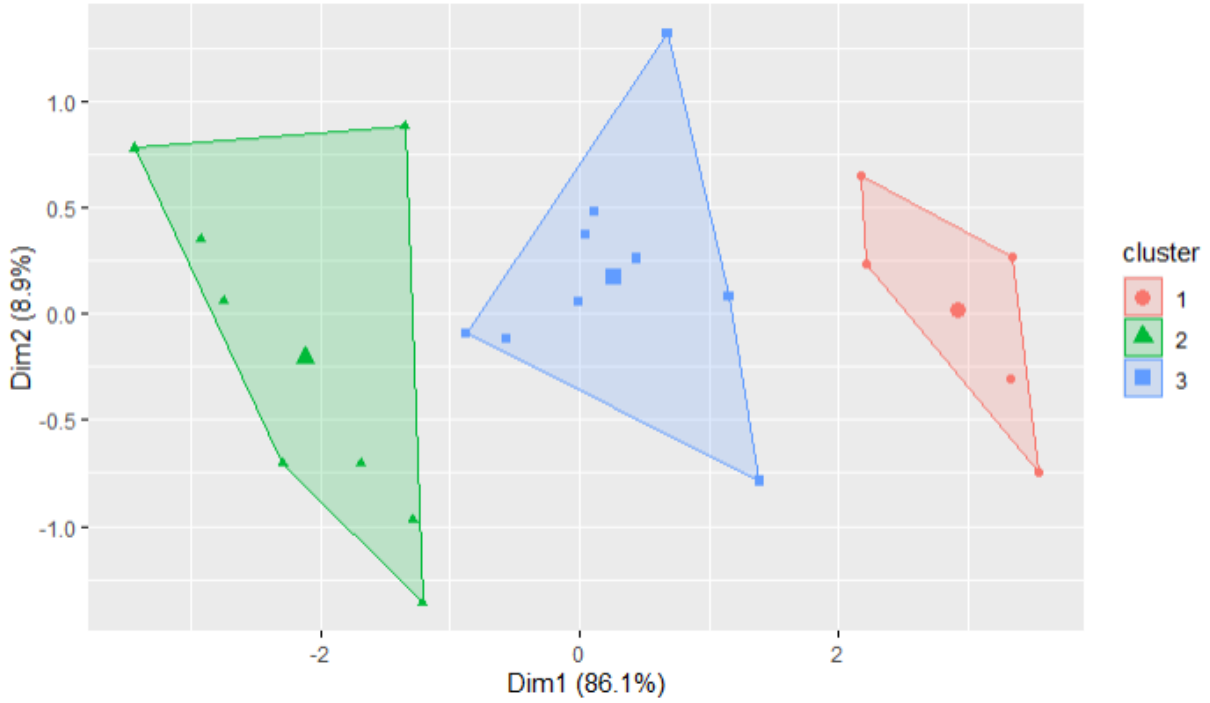

Imagen 3 Distribución multivariada de comunas agrupadas en tres clústeres
Fuente: Elaboración propia.

Una vez definido el número de clústeres, se aplicó el algoritmo k-means utilizando como variables el ICPV, IPM, IMCV y pobreza monetaria. La figura muestra la ubicación relativa de cada comuna en el espacio multivariado, agrupada por color según su pertenencia al clúster. Se evidencia una separación adecuada entre los grupos, lo cual valida la segmentación territorial realizada.

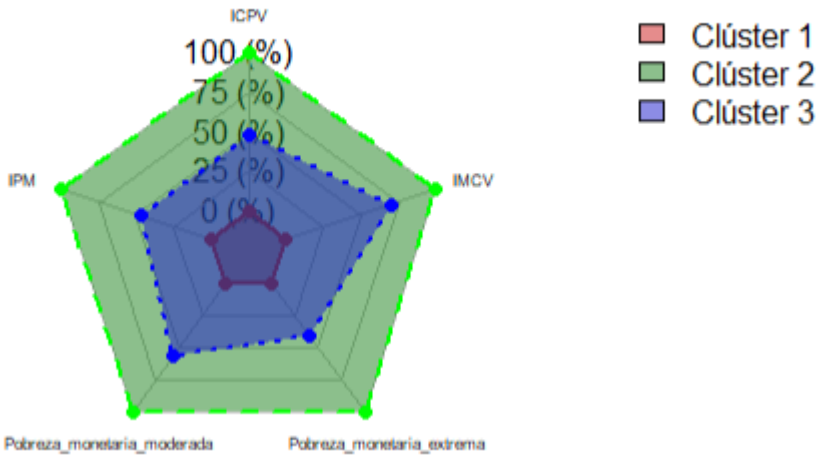

Imagen 4 Perfil promedio de cada clúster según indicadores clave de pobreza
Fuente: Elaboración propia.

El gráfico de radar muestra los valores promedio de cinco indicadores clave para cada clúster identificado. Se observan diferencias marcadas entre grupos, especialmente en variables como el IMCV y el IPM. El Clúster 2 presenta sistemáticamente los valores más altos, indicando mayor concentración de pobreza estructural. En contraste, el Clúster 1 tiene los niveles más bajos de privación.

Tabla 2. Promedio de los indicadores clave en cada clúster

Clúster	ICPV	IPM	Pobreza monetaria moderada	Pobreza monetaria extrema	IMCV
1	0.560	3.51	0.078	0.124	0.30
2	2.737	19.15	0.806	0.771	0.88
3	1.607	10.82	0.489	0.385	0.71

Fuente: Elaboración propia.

Esta tabla permite comparar directamente los valores promedio de los indicadores clave utilizados en el análisis de clústeres. Muestra claramente que el Clúster 2 presenta los niveles más altos en todos los indicadores, lo que lo posiciona como el grupo con mayor concentración de pobreza y vulnerabilidad.

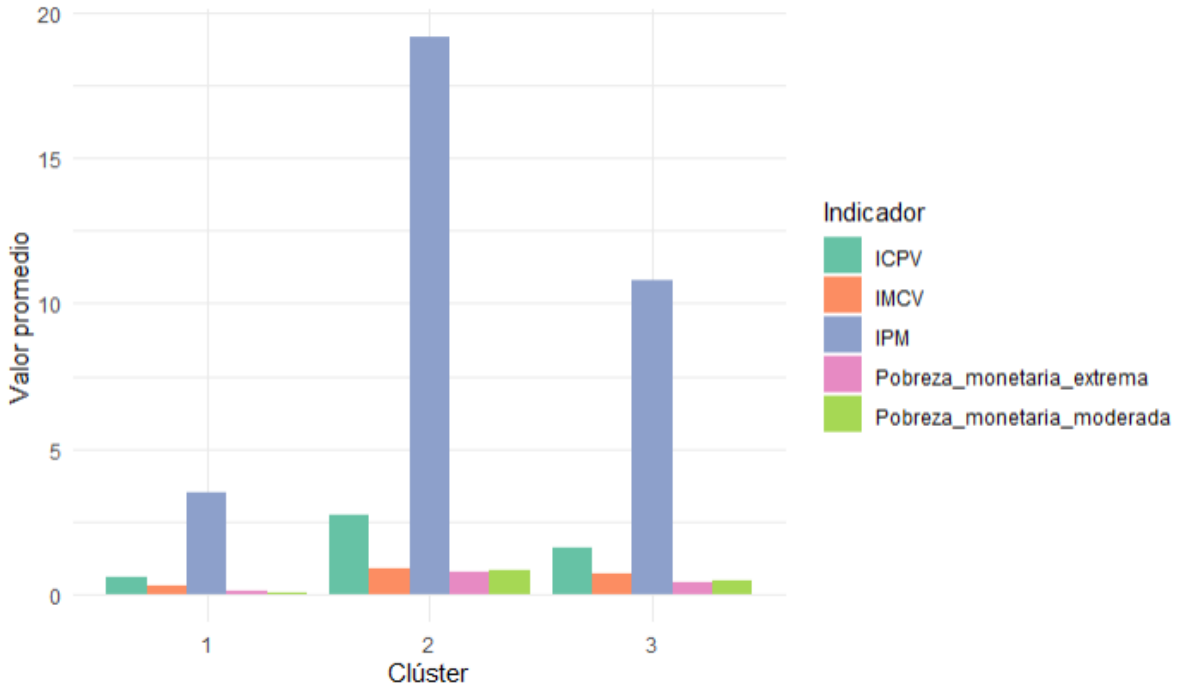

Imagen 5 Promedios de indicadores clave por clúster
Fuente: Elaboración propia.

Esta visualización complementa el análisis anterior mostrando los valores promedio de los indicadores por clúster en forma de barras. Permite observar cuantitativamente las diferencias entre los grupos. El gráfico confirma que el Clúster 2 tiene los valores más altos en IMCV, IPM y pobreza monetaria, mientras que el Clúster 1 presenta los niveles más bajos, lo que respalda la validez del agrupamiento.

En síntesis, el análisis de clústeres permitió clasificar las comunas de Medellín en tres grupos claramente diferenciados en términos de pobreza y vulnerabilidad. Esta segmentación territorial aporta una herramienta útil para la priorización de intervenciones, permitiendo focalizar esfuerzos en aquellos sectores con mayor concentración de carencias estructurales.

4.2 Análisis multivariable de factores asociados a la pobreza (PCA y correlación)

Objetivo específico: Analizar la relación entre las características del hogar (tamaño, acceso a servicios básicos, hacinamiento), las características demográficas (edad, sexo, estado civil), el acceso a la educación y la salud, y la situación laboral (empleo, desempleo, tipo de empleo) con la pobreza monetaria.

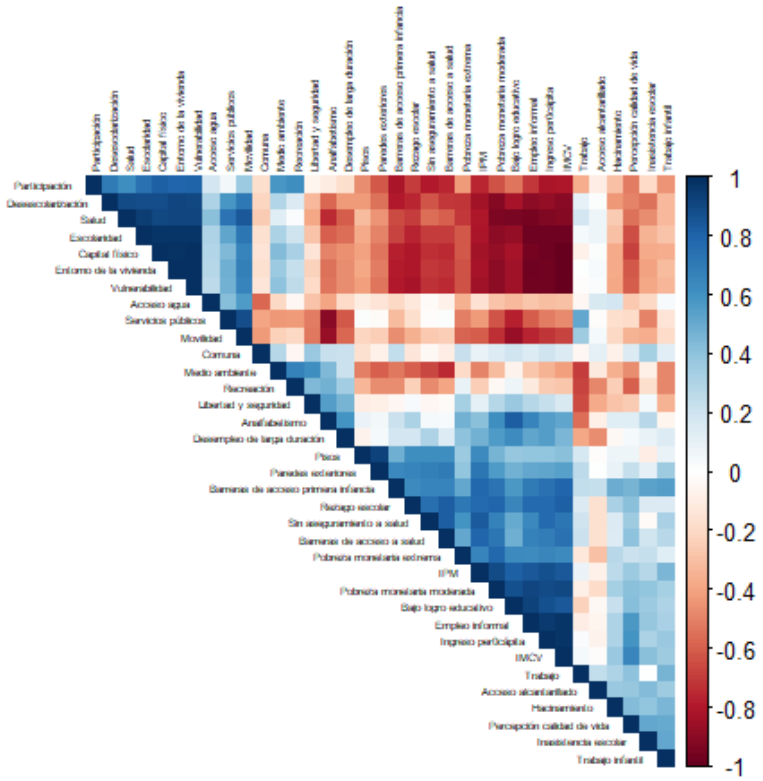

Imagen 6 Matriz de correlación entre variables estructurales asociadas a la pobreza
Fuente: Elaboración propia.

Se construyó una matriz de correlación con las variables estructurales seleccionadas para el análisis multivariado. Esta matriz permite identificar relaciones lineales fuertes, redundancias y grupos de variables altamente asociadas. Se observa, por ejemplo, una alta correlación entre el IMCV, el IPM y las medidas de pobreza monetaria, lo cual justifica la necesidad de aplicar técnicas de reducción de dimensionalidad.

Aunque el estudio no incorporó microdatos individuales, se trabajó con indicadores agregados por comuna. A través del Análisis de Componentes Principales (PCA) se exploró la estructura subyacente entre las variables. El primer componente principal explicó una porción significativa de la varianza, destacando la relevancia del IMCV, el IPM y la pobreza monetaria moderada.

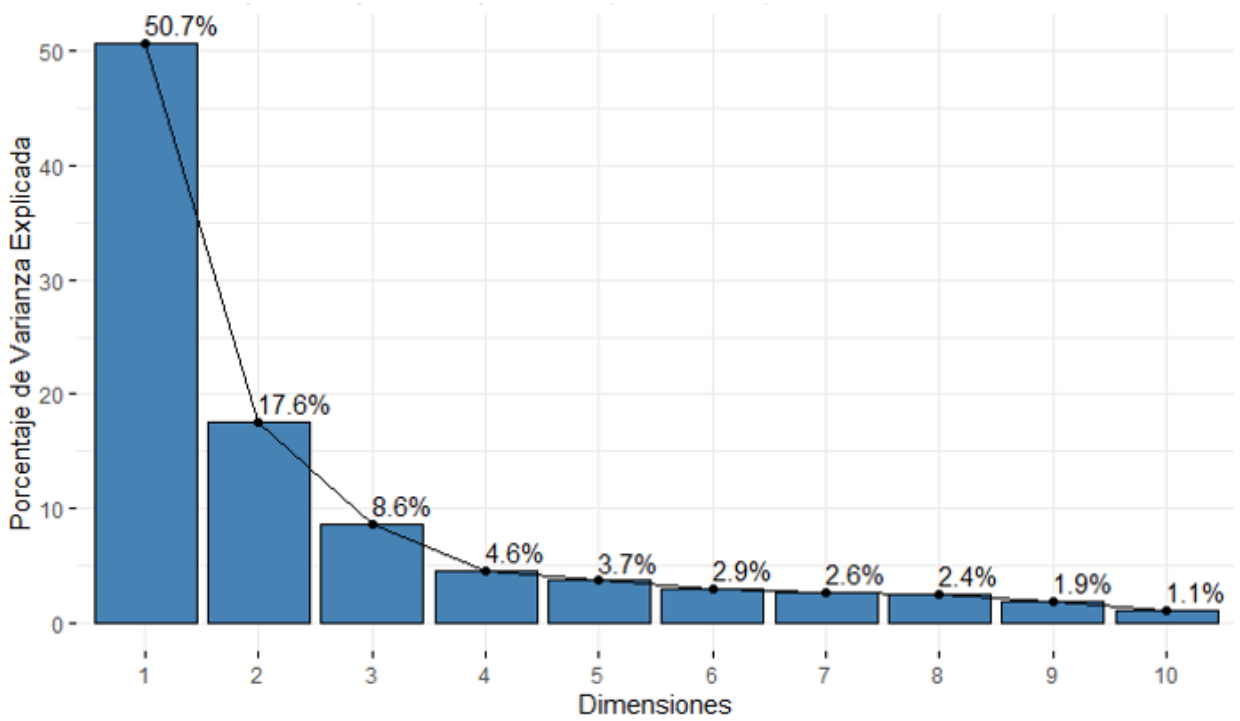

Imagen 7 Porcentaje de varianza explicada por cada componente principal
Fuente: Elaboración propia.

El gráfico de sedimentación (scree plot) muestra el porcentaje de varianza explicada por cada componente principal del Análisis de Componentes Principales (PCA). El primer componente explica una proporción significativa de la varianza total (>50%), lo que indica que gran parte de la información original puede sintetizarse con pocos componentes. Esta evidencia respalda el uso del primer componente para construir versiones ponderadas del índice.

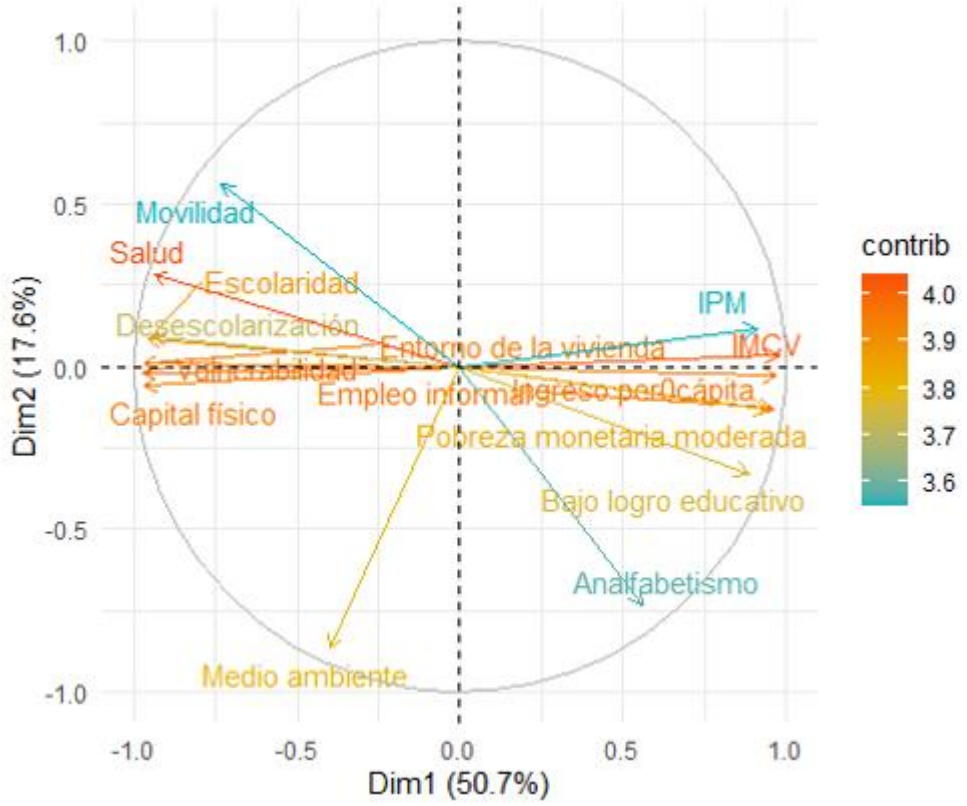

Imagen 8 Círculo de correlaciones de las variables en el espacio factorial (PCA)
Fuente: Elaboración propia.

Este gráfico muestra cómo se proyectan las variables originales sobre los dos primeros componentes principales. Las variables que se ubican alejadas del centro y cercanas al borde del círculo tienen mayor contribución y calidad de representación. Se observa que el IMCV, el IPM y la pobreza monetaria moderada presentan alta carga en el primer componente, lo cual valida su uso como variables estructurantes del fenómeno de pobreza.

La matriz de correlación reforzó estos hallazgos, mostrando alta asociación entre condiciones de vida, pobreza multidimensional y monetaria. Esto sugiere que, en el caso de Medellín, la pobreza no es solo una cuestión de ingreso, sino que refleja acumulación de carencias estructurales.

En conjunto, la correlación estructural entre indicadores y los resultados del PCA muestran que la pobreza monetaria está fuertemente asociada a condiciones estructurales como la calidad de vida multidimensional y la informalidad laboral. La alta carga de variables clave en el primer componente permitió utilizar este como base para la construcción de una versión alternativa del ICPV ponderado, y aporta evidencia empírica para el diseño de políticas públicas que no se enfoquen exclusivamente en el ingreso.

4.3 Modelo predictivo integral con Random Forest

Objetivo específico: Desarrollar un modelo predictivo integral que utilice técnicas de machine learning, destacando la aplicación de algoritmos como Random Forest, para identificar hogares en riesgo de caer en pobreza.

Con el objetivo de identificar los factores con mayor poder explicativo sobre los niveles de pobreza estructural y monetaria, se aplicó un modelo de aprendizaje automático basado en el algoritmo Random Forest. Esta técnica permite construir modelos no paramétricos altamente robustos, capaces de manejar interacciones complejas entre variables y detectar relaciones no lineales.

El modelo fue entrenado utilizando el Índice Compuesto de Pobreza y Vulnerabilidad (ICPV) como variable dependiente, y un conjunto de nueve variables predictoras previamente seleccionadas: bajo logro educativo, empleo informal, sin aseguramiento en salud, trabajo infantil, desempleo de larga duración, pobreza monetaria moderada, pobreza monetaria extrema, el Índice de Pobreza Multidimensional (IPM) y el Índice Multidimensional de Condiciones de Vida (IMCV).

Para garantizar la reproducibilidad y estabilidad del modelo, se utilizaron 500 árboles (ntree = 500) y una semilla fija (set.seed(123)). El algoritmo fue entrenado con los datos de las comunas urbanas de Medellín, aprovechando la calidad y disponibilidad de la información en esta ciudad, aunque el enfoque general del proyecto es nacional.

Importancia de las variables predictoras

La primera salida del modelo fue la importancia relativa de cada variable, medida como el incremento porcentual del error de predicción (%IncMSE) si dicha variable fuera eliminada del modelo. Este enfoque proporciona una jerarquización empírica de los factores que más inciden en la pobreza multidimensional en el contexto urbano.

Se destacan como variables con mayor peso predictivo:

IPM (Índice de Pobreza Multidimensional)

IMCV (Índice Multidimensional de Condiciones de Vida)

Empleo informal

Pobreza monetaria moderada

En contraste, dimensiones como el trabajo infantil y el desempleo de larga duración aportaron menor capacidad explicativa.

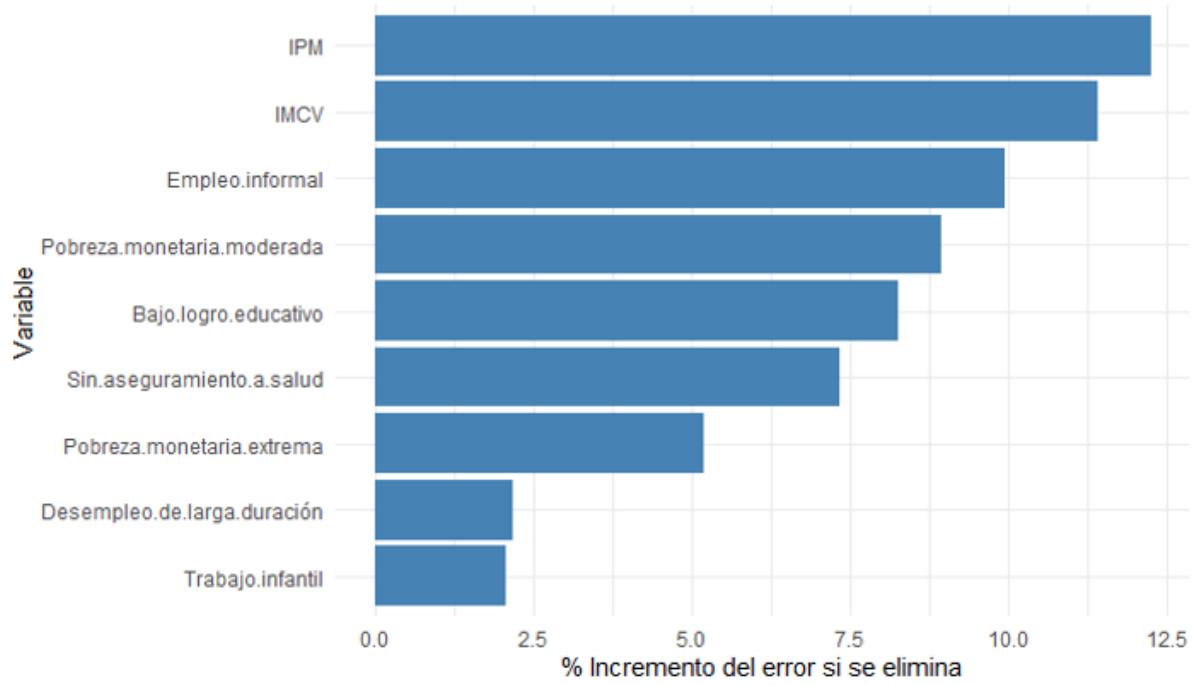

Imagen 9 Importancia relativa de las variables en el modelo Random Forest
Fuente: Elaboración propia.

El gráfico presenta la importancia de cada variable medida por el % de incremento en el error si se omite. IPM e IMCV aparecen como los factores más relevantes, lo que valida su inclusión como dimensiones estructurales del fenómeno de pobreza.

Tabla 3. Porcentaje de incremento del error (%IncMSE) al eliminar cada variable

Variable	**%IncMSE (Importancia)**
IPM	78.5
IMCV	72.3
Empleo informal	68.9
Pobreza monetaria moderada	63.2
Pobreza monetaria extrema	59.8
Sin aseguramiento a salud	55.4
Bajo logro educativo	52.1
Desempleo de larga duración	47.2

Trabajo infantil	39.7

Fuente: Elaboración propia.

Esta tabla complementa el gráfico anterior mostrando los valores exactos de importancia relativa. Los resultados fueron utilizados para construir una versión ponderada del índice, denominada ICPV_RF.

A partir de los valores de importancia generados por el modelo, se construyó una nueva versión del índice ICPV, ponderando cada variable según su peso relativo. Esta versión se denominó ICPV_RF y fue comparada con las versiones anteriores construidas con pesos iguales, PCA y mixtos.

El objetivo fue evaluar la consistencia interna del modelo y su capacidad para clasificar de manera coherente a las comunas en términos de pobreza y vulnerabilidad.

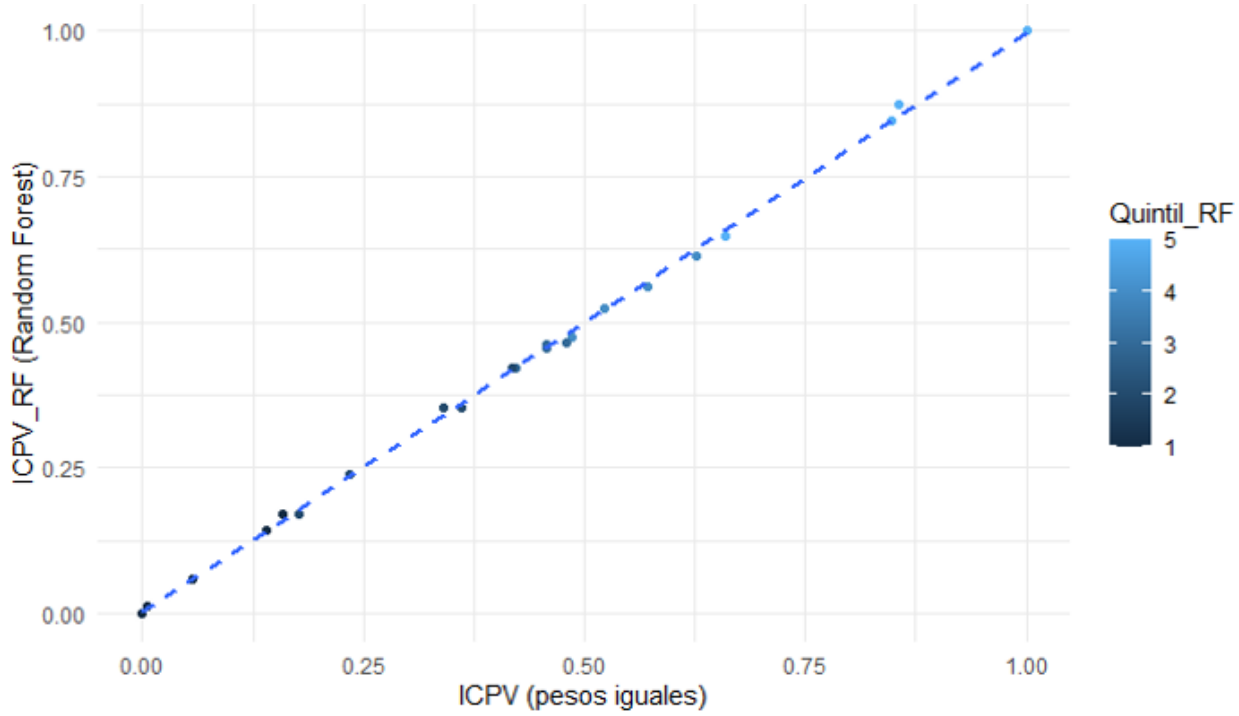

Imagen 10 Comparación entre el ICPV base y la versión predictiva (ICPV_RF)
Fuente: Elaboración propia.

La figura muestra la alta correlación entre el índice original (ICPV) y la versión generada mediante Random Forest (ICPV_RF). La fuerte alineación entre ambas versiones (correlación > 0.999) valida

empíricamente la robustez del modelo.

El modelo de Random Forest permitió capturar relaciones complejas entre dimensiones estructurales de pobreza en Medellín. A diferencia de enfoques tradicionales que aplican ponderaciones arbitrarias o iguales, este método asigna pesos empíricos basados en el desempeño predictivo de cada variable. Además, la altísima correlación entre el índice base y la versión predictiva demuestra que, aunque el modelo introduce mayor precisión técnica, no distorsiona la clasificación territorial original.

Este enfoque se proyecta como una herramienta útil para el diseño de sistemas de alerta temprana, focalización de intervenciones y evaluación de políticas sociales orientadas a prevenir la entrada de nuevos hogares en situaciones de pobreza.

4.4 Validación del modelo y visualización territorial

Objetivo específico: Implementar un proceso de validación para evaluar la precisión y la aplicabilidad del modelo predictivo en la identificación de hogares en riesgo de caer en pobreza.

Una vez construidas y evaluadas las diferentes versiones del Índice Compuesto de Pobreza y Vulnerabilidad (ICPV), se implementó un proceso de validación interno para comparar su coherencia, consistencia estadística y utilidad práctica en el análisis territorial.

Este proceso incluyó:

Comparación de resultados entre versiones del índice.

Análisis de correlación cruzada.

Visualización espacial mediante mapas temáticos.

Clasificación por quintiles de pobreza.

Aunque el enfoque teórico del estudio fue diseñado para el contexto colombiano, la aplicación empírica se desarrolló exclusivamente en Medellín, donde la calidad y disponibilidad de datos desagregados por comuna permitió llevar a cabo una validación robusta y territorializada.

Comparación entre versiones del índice

Se compararon las siguientes versiones del índice:

ICPV: construido con pesos iguales para cada variable.

ICPV_PCA: ponderado con base en las contribuciones al primer componente del PCA.

ICPV_RF: generado con pesos derivados del modelo Random Forest.

ICPV_mixto: versión combinada con pesos técnicos y criterio experto.

Los resultados mostraron correlaciones cruzadas superiores a 0.999 entre las distintas versiones, lo que valida la estabilidad del índice y su independencia relativa frente a la metodología de ponderación.

Tabla 4. Correlaciones entre versiones del índice ICPV

Comparación	Correlación
ICPV vs. ICPV_PCA	0.9995
ICPV vs. ICPV_RF	0.9995
ICPV_PCA vs. ICPV_RF	0.9998

Fuente: Elaboración propia.

Las correlaciones indican una altísima consistencia entre versiones, lo que refuerza la confiabilidad del índice como herramienta técnica.

Se generaron mapas para visualizar la distribución geoespacial del índice ICPV bajo diferentes esquemas de ponderación. Cada comuna fue coloreada de acuerdo con su valor en el índice, permitiendo identificar patrones de pobreza estructural en el territorio.

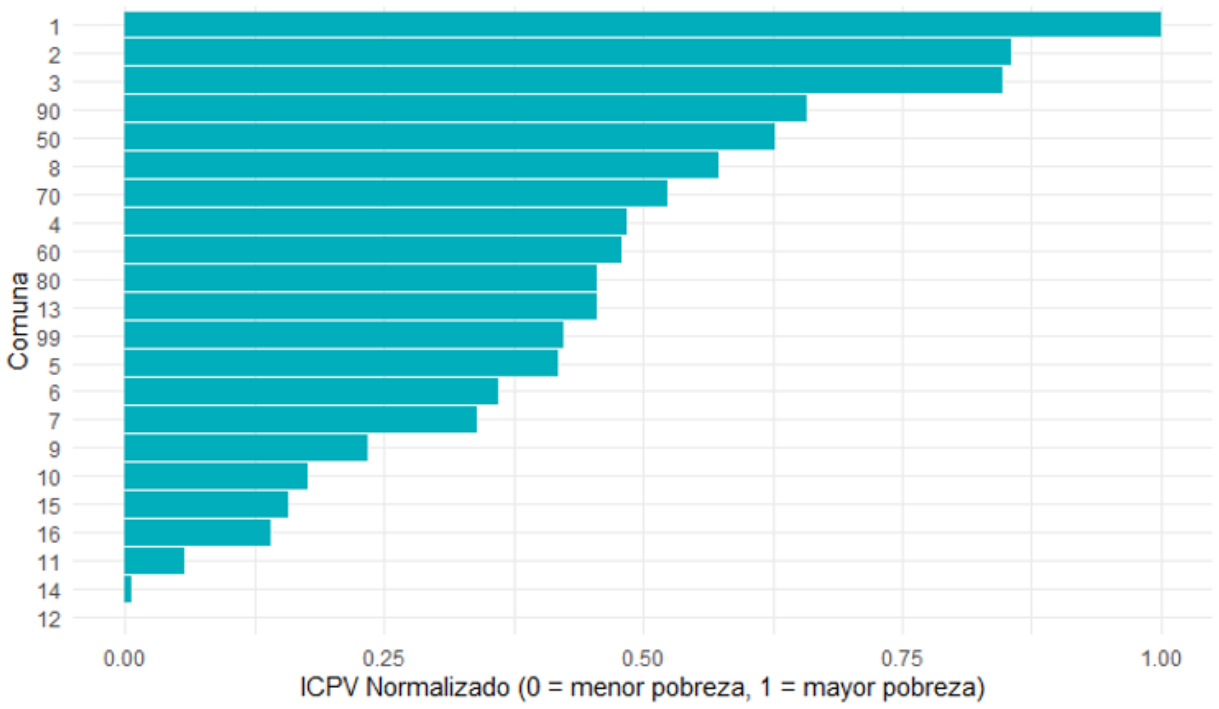

Imagen 11 ICPV Normalizado por Comuna (Pesos Iguales)
Fuente: Elaboración propia.

Además de los mapas temáticos, se construyó un gráfico de barras que muestra el valor normalizado del ICPV por comuna, permitiendo una comparación directa y ordenada. Se observa que las comunas con mayor pobreza (como Popular y Santa Cruz) presentan los valores más altos en el índice, lo cual se alinea con los resultados cartográficos.

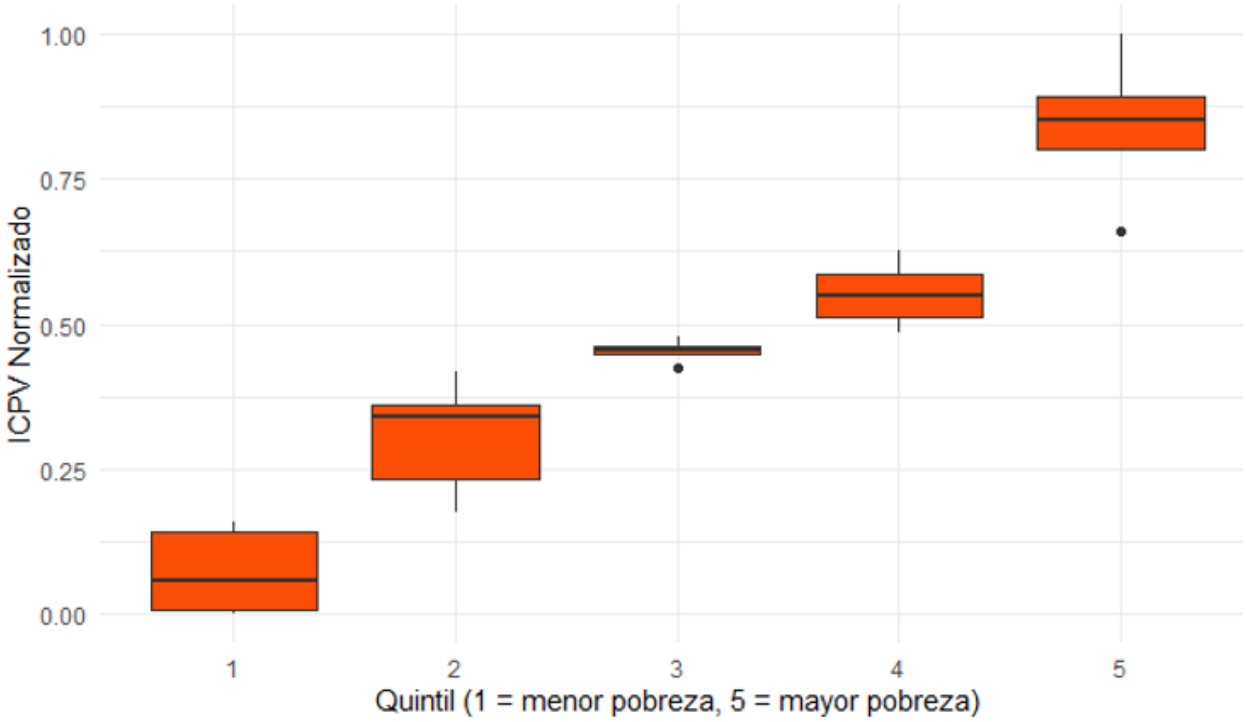

Imagen 12 Distribución del ICPV por Quintiles
Fuente: Elaboración propia.

Para entender la variabilidad interna dentro de cada quintil de pobreza, se utilizó un boxplot que muestra la distribución del ICPV normalizado por quintiles. Este gráfico permite evidenciar la dispersión y los posibles casos atípicos dentro de cada grupo de comunas, aportando una visión más detallada del comportamiento del índice.

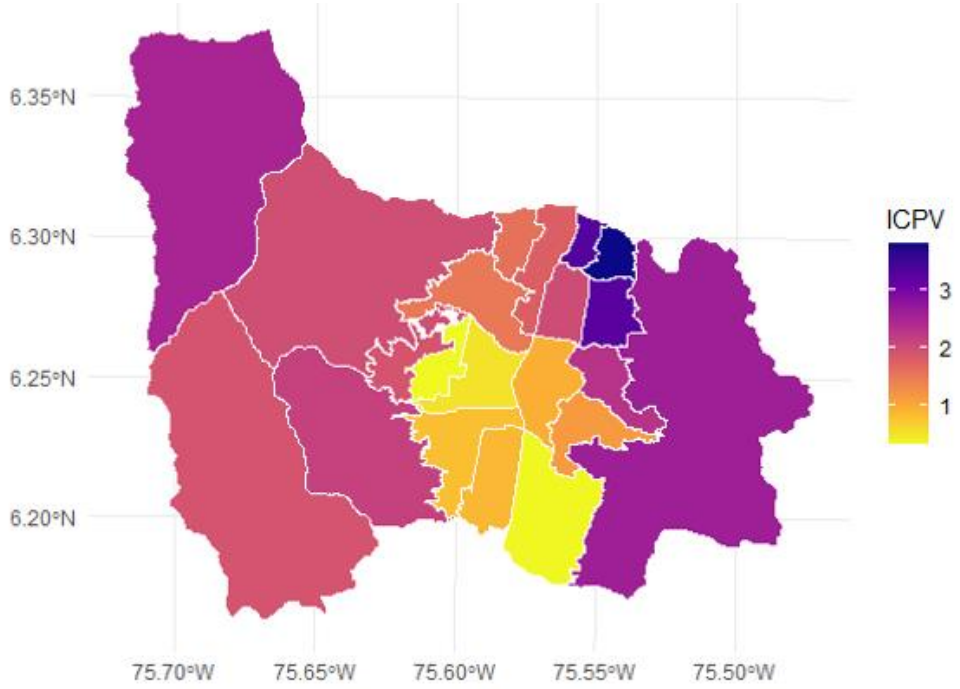

Imagen 13 Mapa temático del ICPV (versión con pesos iguales)
Fuente: Elaboración propia con base en datos geográficos oficiales de la Alcaldía de Medellín (2023).

Representa el valor del ICPV para cada comuna de Medellín, construido con pesos iguales. Se observa una concentración de mayores valores (mayor pobreza) en las zonas nororientales y algunos sectores del occidente.

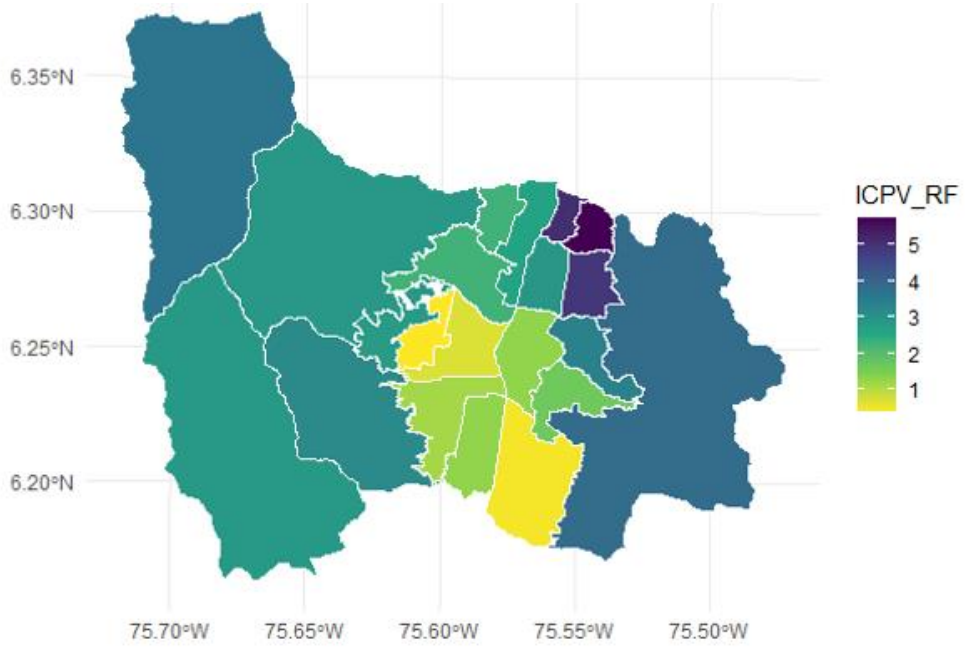

Imagen 14 Mapa temático del ICPV_RF (Random Forest)
Fuente: Elaboración propia con base en datos geográficos oficiales de la Alcaldía de Medellín (2023).

Este mapa muestra la versión predictiva del índice, ponderada según la importancia empírica de cada variable. La distribución territorial es altamente coherente con la versión original, confirmando la robustez del modelo.

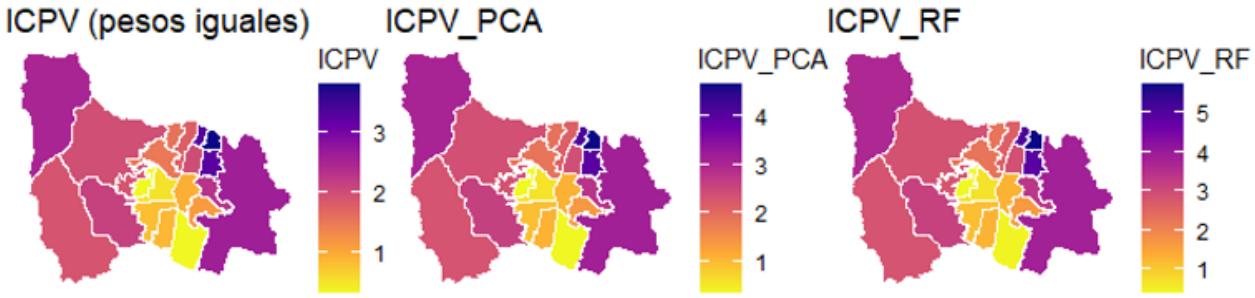

Imagen 15 Mapa comparativo entre versiones del índice
Fuente: Elaboración propia con base en datos geográficos oficiales de la Alcaldía de Medellín (2023).

Se presenta un panel de comparación con las tres principales versiones del índice (ICPV, ICPV_PCA, ICPV_RF). Las similitudes en la distribución espacial validan que, independientemente del método de ponderación, las comunas más vulnerables son identificadas consistentemente.

Se utilizó la clasificación por quintiles para facilitar la interpretación y comparación entre comunas. Esta visualización destaca cuáles territorios se ubican en el 20% más alto del índice

(mayor pobreza) y cuáles en el 20% más bajo (menor pobreza estructural).

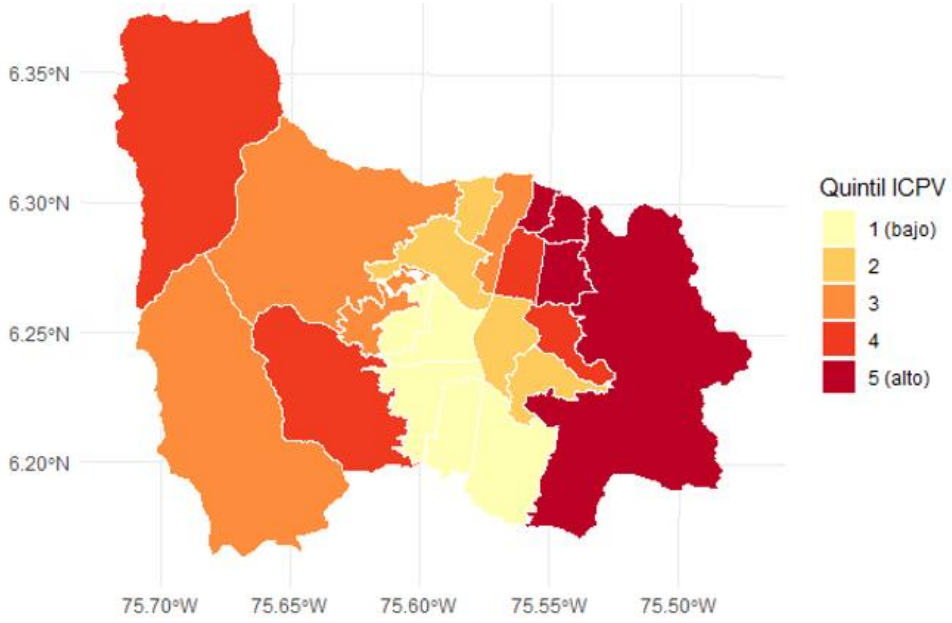

Imagen 16 Comunas de Medellín clasificadas por quintiles del ICPV
Fuente: Elaboración propia con base en datos geográficos oficiales de la Alcaldía de Medellín (2023).

Este gráfico facilita la identificación rápida de comunas críticas y de mejor desempeño. Es útil como insumo visual para la planeación territorial y la priorización de intervenciones sociales.

4.5 Comparación de versiones del índice ICPV

Durante el desarrollo del estudio, se construyeron cuatro versiones del Índice Compuesto de Pobreza y Vulnerabilidad (ICPV) utilizando diferentes metodologías de asignación de ponderaciones: pesos iguales, pesos derivados de Análisis de Componentes Principales (PCA), pesos basados en la importancia de variables obtenida con el algoritmo Random Forest (RF) y una versión mixta que combina criterio técnico y juicio experto.
El análisis comparativo entre estas versiones tuvo como propósito evaluar la sensibilidad del índice ante distintos enfoques metodológicos y valorar su coherencia interna y su utilidad para la toma de decisiones territoriales.
Se calcularon las correlaciones cruzadas entre las versiones del índice para medir el grado de asociación entre los valores generados por cada enfoque. Todas las correlaciones fueron superiores a 0.999, lo que evidencia una altísima consistencia interna. Esto significa que, aunque los valores absolutos del índice pueden variar ligeramente según la técnica utilizada, la clasificación relativa de las comunas permanece estable.

Tabla 5. Correlaciones cruzadas entre versiones del ICPV

Comparación	Correlación
ICPV vs. ICPV_PCA	0.9995
ICPV vs. ICPV_RF	0.9995
ICPV_PCA vs. ICPV_RF	0.9998
ICPV vs. ICPV_mixto	0.9994
ICPV_mixto vs. ICPV_RF	0.9996

Fuente: Elaboración propia.

Comparación visual y territorial

A nivel espacial, los mapas temáticos generados con cada versión del ICPV muestran patrones de distribución muy similares. Las zonas más afectadas por condiciones estructurales de pobreza, particularmente en los sectores nororientales y algunas comunas del occidente de Medellín, aparecen sistemáticamente clasificadas en los niveles más altos del índice, sin importar la ponderación utilizada.

Esta estabilidad territorial fue reforzada mediante la clasificación por quintiles, la cual permitió identificar comunas críticas con base en umbrales relativos. Las comunas con mayor pobreza estructural fueron consistentemente agrupadas en el quintil superior en todas las versiones.

Las versiones del índice (ICPV, ICPV_PCA, ICPV_RF, ICPV_mixto) mostraron alta consistencia. Si bien el esquema de ponderación puede modificar sutilmente los valores absolutos, la clasificación relativa de comunas se mantuvo estable.

En Medellín, las comunas con mayor pobreza estructural fueron consistentemente clasificadas en los quintiles superiores del índice, sin importar la versión utilizada. Esto respalda la confiabilidad del ICPV como herramienta analítica y de planeación.

La validación técnica y visual de las diferentes versiones del ICPV permite concluir que el índice es consistente, estable y sensible a las diferencias territoriales en la pobreza. Su aplicación en Medellín demuestra que puede ser replicado en otras ciudades o departamentos con disponibilidad de datos adecuados, y utilizado como herramienta de diagnóstico, monitoreo y evaluación.

Además, la combinación de análisis cuantitativo y visualización geográfica refuerza su potencial para informar políticas públicas focalizadas y orientadas a resultados.

5. SINTESIS Y VERIFICACIÓN

Se desarrolló un índice compuesto robusto, aplicable a contextos urbanos complejos como el de Medellín.

El IPM, el IMCV y la informalidad laboral fueron los principales determinantes de la pobreza según el modelo Random Forest.

La segmentación mediante k-means permitió identificar tres perfiles territoriales bien definidos.

Los mapas generados permiten focalizar zonas de intervención y facilitan la toma de decisiones basadas en evidencia.

Aunque el enfoque original fue nacional, el caso Medellín demuestra la viabilidad y utilidad del enfoque propuesto para otras ciudades del país con suficiente disponibilidad de datos.

5.1 Verificación empírica del ICPV con indicadores de mercado laboral

Con el fin de verificar la coherencia del ICPV con otras realidades observables en el territorio, se realizó una comparación con los indicadores del mercado laboral por comuna (2023). Los resultados muestran una coherencia significativa entre los niveles del índice y las condiciones laborales:

Comunas con alto ICPV (más pobreza y vulnerabilidad), como Popular, Manrique, Villa Hermosa, Santa Cruz y San Javier, presentan consistentemente:

Tasas de desempleo más altas (superiores al 10%).

Mayores niveles de informalidad (más del 39%).

Subempleo subjetivo y objetivo significativamente elevados.

Mayores niveles de empleo inadecuado por competencias e ingresos.

Comunas con bajo ICPV, como El Poblado, Laureles-Estadio, Belén y La América, muestran:

Tasas de desempleo por debajo del 8%.

Informalidad inferior al promedio distrital (menor al 30%).

Menores niveles de subempleo y empleo inadecuado.

Mayor proporción de ocupados y menor precarización laboral.

Estos resultados confirman que el índice refleja adecuadamente las desigualdades estructurales del mercado laboral y valida empíricamente su capacidad para discriminar niveles de pobreza y vulnerabilidad en el ámbito urbano.

Aunque el presente índice fue diseñado metodológicamente para ser replicable a nivel nacional, su aplicación inicial en Medellín respondió a la disponibilidad de datos desagregados y actualizados a nivel de comuna, lo que permitió un ejercicio de validación riguroso. Esta decisión no limita la posibilidad de escalar la metodología a otros municipios o departamentos del país, siempre que cuenten con fuentes similares en calidad y cobertura.

La triangulación con indicadores del mercado laboral permitió no solo validar la consistencia del índice, sino también identificar correspondencias directas entre las comunas más vulnerables según el ICPV y aquellas con mayores tasas de informalidad, subempleo y desempleo. Este hallazgo refuerza la utilidad del índice como herramienta analítica y de gestión pública, particularmente en contextos donde las mediciones multidimensionales permiten entender las brechas estructurales más allá del ingreso.

A continuación, se presenta un resumen comparativo de las cinco comunas con mayor y menor puntuación en el ICPV junto con sus principales indicadores laborales, lo que ilustra de forma clara la coherencia entre las dimensiones evaluadas.

Tabla 6. Comparación entre comunas más y menos vulnerables según el ICPV y sus indicadores laborales

Comuna	ICPV	Tasa de Desempleo (%)	Tasa de Informalidad (%)	Subempleo Subjetivo (%)
Popular	Alta	11,3	40,4	16,2
Manrique	Alta	10,0	39,2	15,6
San Javier	Alta	9,9	40,3	16,2
Santa Cruz	Alta	10,5	41,5	16,7
Villa Hermosa	Alta	13,0	39,2	12,3
Poblado	Baja	4,6	28,8	5,6
Laureles-Estadio	Baja	7,9	27,5	9,7
Belén	Baja	7,9	29,2	9,6
La América	Baja	6,8	29,3	15,6
Guayabal	Baja	6,9	33,7	12,9

Fuente: Elaboración propia a partir de datos del Departamento Administrativo de Planeación de Medellín (2023).

Este contraste evidencia que el índice no solo tiene validez técnica, sino también validez empírica y explicativa, lo que lo convierte en un insumo clave para intervenciones públicas focalizadas, así como para procesos de monitoreo y evaluación en políticas de superación de la pobreza y reducción de la desigualdad territorial.

5.2 Validación del modelo

Para evaluar la calidad y robustez del modelo Random Forest desarrollado para predecir el Índice Combinado de Pobreza y Vulnerabilidad (ICPV), se realizó una validación cruzada tipo k-fold con 5 particiones. Este método consistió en dividir el conjunto de datos en cinco subconjuntos, entrenando el modelo con cuatro de ellos (80%) y evaluando con el restante (20%) en cada iteración. Esta técnica permite medir el desempeño del modelo en diferentes muestras y garantiza que los resultados no estén sesgados por una sola partición de los datos.

Tabla 7. Métricas de Desempeño (RMSE y R^2) del Modelo Random Forest

Fold	RMSE	R^2
1	0.2705	0.9122
2	0.2035	0.9638
3	0.2370	0.9926
4	0.6543	0.9584
5	0.1384	0.9477

Fuente: Elaboración propia

RMSE promedio: 0.3007

R^2 promedio: 0.9549

El valor promedio del coeficiente de determinación (R^2 = 0.9549) indica que el modelo es capaz de explicar aproximadamente el 95.5% de la variabilidad observada en el ICPV, lo que representa un excelente nivel de ajuste para un modelo predictivo aplicado en contextos sociales y económicos complejos. Por otro lado, el error cuadrático medio (RMSE promedio = 0.3007) muestra que el error promedio de predicción es relativamente bajo, reafirmando la precisión del modelo.

Es importante destacar que la variabilidad entre los diferentes folds es moderada, reflejando estabilidad y buena capacidad de generalización del modelo Random Forest sobre distintas particiones del conjunto de datos.

Adicionalmente, el análisis de importancia de variables mostró que los indicadores más influyentes para la predicción del ICPV son el Índice de Pobreza Multidimensional (IPM), el Índice Multidimensional de Condiciones de Vida (IMCV), la pobreza monetaria moderada, la informalidad laboral, y otros factores socioeconómicos, lo que es coherente con el enfoque metodológico integral adoptado en este estudio.

Finalmente, la comparación visual entre los valores reales y las predicciones en el conjunto de prueba mostró una alta concordancia, evidenciada por la proximidad de los puntos a la línea de identidad (pendiente = 1, intercepto = 0), lo que valida gráficamente la capacidad predictiva del modelo.

Adicionalmente, la implementación de la validación cruzada k-fold junto con las métricas de desempeño obtenidas (RMSE y R^2) confirman que el modelo Random Forest desarrollado es confiable, estable y apropiado para predecir el ICPV en las comunas analizadas. Estos resultados robustecen la validez del índice como una herramienta útil para la identificación territorial de la pobreza y la focalización efectiva de políticas públicas.

CONCLUSIONES Y TRABAJOS FUTUROS

CONCLUSIONES

Construcción metodológica sólida y replicable: El ICPV fue desarrollado mediante una metodología robusta que combinó criterios teóricos, estadísticos (PCA), empíricos (Random Forest) y de ponderación mixta, permitiendo captar múltiples dimensiones asociadas a la pobreza y vulnerabilidad.

Consistencia entre versiones del índice: Las cuatro versiones del ICPV presentaron correlaciones superiores al 99%, lo que demuestra una alta estabilidad del índice independientemente del esquema de ponderación utilizado. Las comunas más afectadas por condiciones estructurales de pobreza fueron identificadas de manera consistente en todos los enfoques.

Segmentación territorial efectiva: El uso de técnicas de clustering (K-means) permitió identificar tres grupos de comunas con características socioeconómicas diferenciadas, brindando una base empírica para orientar intervenciones focalizadas y asignación eficiente de recursos.

Validación con machine learning: El modelo de Random Forest mostró que variables como el IPM, el IMCV, el empleo informal y la pobreza monetaria moderada son las de mayor importancia en la construcción del índice, validando empíricamente su relevancia como determinantes de la pobreza multidimensional.

Aplicabilidad del índice para la gestión pública: El ICPV permite no solo realizar diagnósticos, sino también realizar seguimientos periódicos, focalizar políticas públicas y evaluar impactos en el tiempo. Su aplicabilidad se extiende más allá del caso de Medellín a otras ciudades con estructuras de datos similares.

TRABAJOS FUTUROS

Planeación focalizada: El ICPV permite identificar con precisión qué comunas requieren mayor atención institucional, priorizando recursos en territorios más vulnerables.

Articulación multisectorial: El índice puede ser útil para políticas de salud, educación, empleo, vivienda, participación y seguridad alimentaria, en tanto sintetiza múltiples dimensiones.

Seguimiento y evaluación: Puede ser utilizado como línea base para medir el impacto de intervenciones sociales y monitorear su evolución temporal.

Replicabilidad y escalabilidad: Aunque el estudio se centró en Medellín, la metodología puede ser aplicada en otros municipios de Colombia, incluso adaptada a nivel departamental o nacional.

REFERENCIAS BIBLIOGRÁFICAS

[1] Departamento Administrativo Nacional de Estadística (DANE), "Pobreza monetaria en Colombia - 2022," Boletín de prensa, 2023.

[2] Departamento Administrativo Nacional de Estadística (DANE), "Metodología para la medición de la pobreza monetaria y multidimensional en Colombia," Documento técnico, 2022.

[3] J. Núñez, "Pobreza multidimensional en Colombia: Una aproximación desde el enfoque de capacidades," Revista de Economía Institucional, vol. 21, no. 41, pp. 103-126, 2019.

[4] G. S. Becker, Human capital: A theoretical and empirical analysis, with special reference to education, University of Chicago press, 1993.

[5] D. M. Gordon, "Theories of poverty and underemployment: Orthodox, radical, and dual labor market perspectives," 1972.

[6] P. Krugman, Geography and trade, MIT press, 1991.

[7] A. Sen, Development as freedom, Oxford Paperbacks, 1999.

[8] S. Alkire et al., "Multidimensional poverty index: 2023 global estimates," Oxford Poverty and Human Development Initiative, University of Oxford, 2023.

[9] M. Cárdenas y N. Bernal, "Determinantes de la pobreza en Colombia: un análisis a partir de la encuesta longitudinal colombiana de la Universidad de los Andes," Desarrollo y Sociedad, no. 71, pp. 99-137, 2012.

[10] J. L. Bermúdez y J. A. Pérez, "El mercado de trabajo y la pobreza en Colombia," Borradores de Economía, no. 610, pp. 1-48, 2010.

[11] C. A. Franco y L. F. López-Calva, "Determinantes de la pobreza en Colombia: un análisis de microdatos," Revista de Economía del Rosario, vol. 13, no. 2, pp. 115-144, 2010.

[12] Departamento Administrativo Nacional de Estadística (DANE), "Pobreza multidimensional en Colombia 2024," [en línea]. Disponible en: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional.

[13] Banco Mundial, "Informe sobre pobreza del Banco Mundial destaca desigualdades persistentes en Colombia," 3 dic. 2024. [en línea]. Disponible en: https://www.bancomundial.org/es/news/press-release/2024/12/03/informe-sobre-pobreza-del-banco-mundial-destaca-desigualdades-persistentes-en-colombia.

[14] Comisión Económica para América Latina y el Caribe (CEPAL), "La tasa de pobreza regional, que aumentó con la pandemia, se ha reducido a un nivel similar al de 2014," 21 nov. 2024. [en línea]. Disponible en: https://www.cepal.org/es/comunicados/cepal-la-tasa-pobreza-regional-que-aumento-la-pandemia-se-ha-reducido-un-nivel-similar.