Clasificación socioeconómica de hogares de Medellín

Proyecto Integrador

INTEGRANTES

- Catalina Piedrahita Jaramillo – Ingeniera Industrial
- Susana Londoño Muñoz - Bióloga
- David Betancur Londoño - Economista
- Diego Andrés Jaramillo Zapata – Ingeniero de sistemas
- Diego Andrés Valderrama Laverde - Ingeniero de sistemas

“ No todo lo que se puede contar cuenta, y no todo lo que cuenta puede ser contado”

—ALBERT EINSTEIN

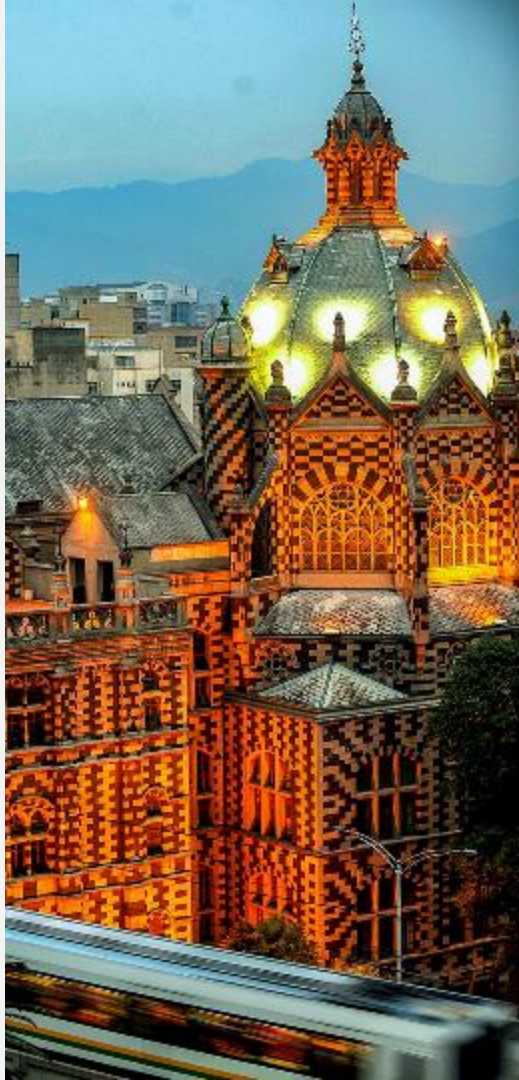
01 JUSTIFICACIÓN

- Identificar grupos poblacionales con características similares es un instrumento para una variedad de objetivos económicos y sociales.

- Una de las clasificaciones más utilizada es el estrato socioeconómico de la vivienda. Se usan para definir clases sociales, cobro de tarifas, costos educativos, acceso a subsidios, acceso a espacios recreativos, campañas comerciales o el cobro de valorización.

- La estratificación representa una metodología de segregación socioespacial, donde se acentúan las diferencias territoriales. Estudios han demostrado que en algunos casos no hay correlación positiva entre la capacidad de pago y estrato.

- *Bajo este contexto, se propone para los hogares de Medellín usar variables socioeconómicas de los individuos, vivienda y variables de entorno para determinar grupos sociales y hacer clasificaciones*.



02 OBJETIVOS

General

Desarrollar para los hogares de Medellín una clasificación que permita conformar grupos similares a nivel socioeconómico a partir de diferentes variables sociodemográficas, incluyendo, algunas de entorno y percepción.

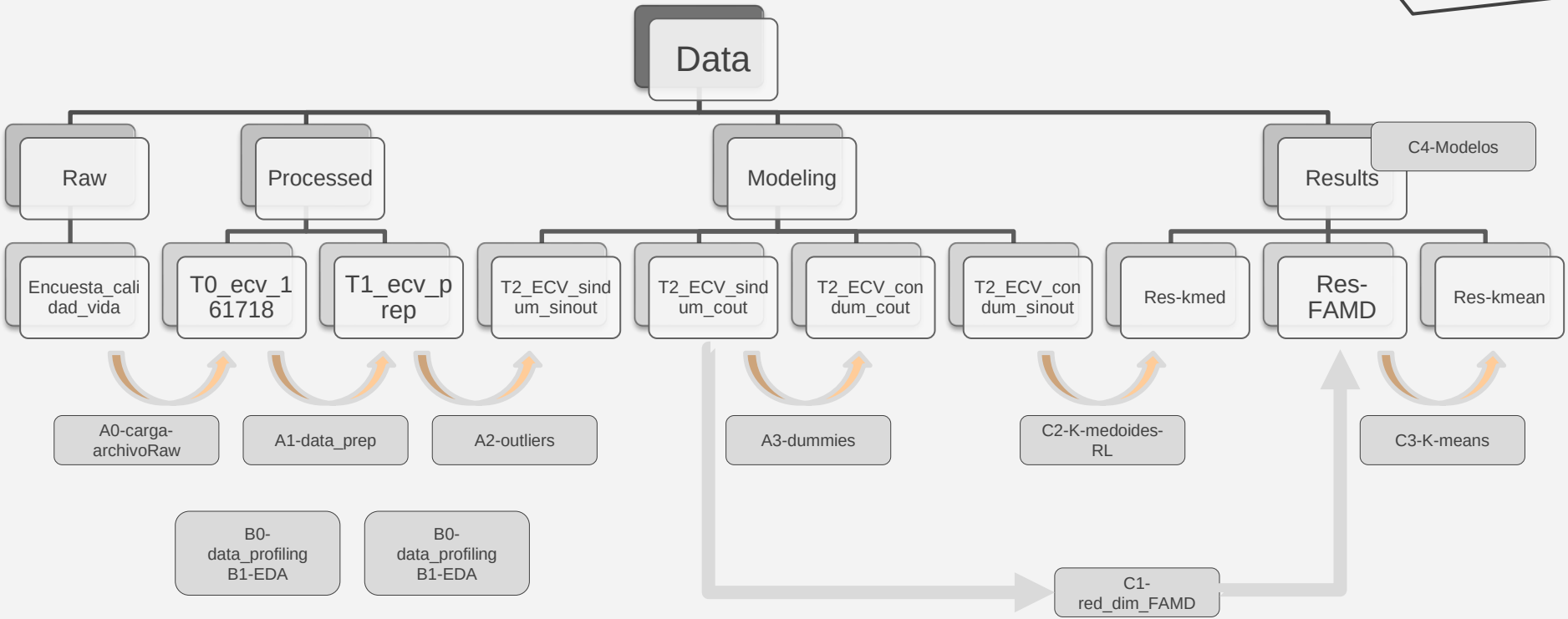
Específicos

- Estudiar información alrededor del tema bajo análisis.
- Preparar los datos para el análisis: exploración, identificación de variables, limpieza de datos, eliminación de outliers, ordenación según marco conceptual.
- Aplicar una metodología para reducir dimensionalidad.
- Probar alternativas de clasificación socioeconómica, con modelos no supervisados y modelos supervisados.
- Concluir sobre la clasificación de los hogares.

03
METODOLOGÍA

Limpieza	Reducción de Dimensionalidad	Modelos no supervisados	Modelos supervisados
• Perfilamiento de variables del dataset • Identificación de outliers: Definición de centro y distancias de Gower	• FAMD	• K-Medoids • K-Means	• Árboles de decisión • Regresión Logística • Naive Bayes

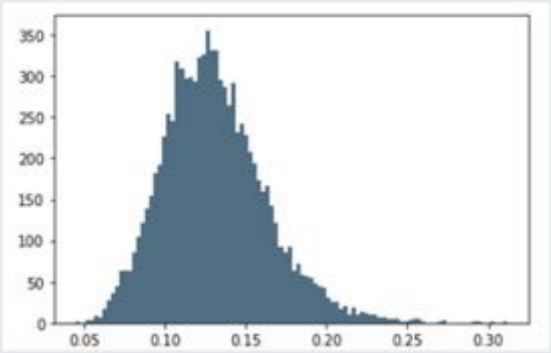
Flujo de datos

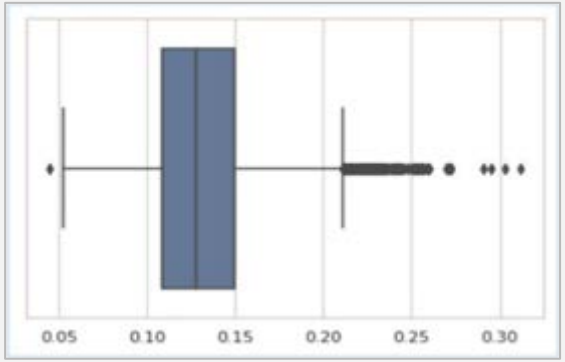


04

RESULTADOS

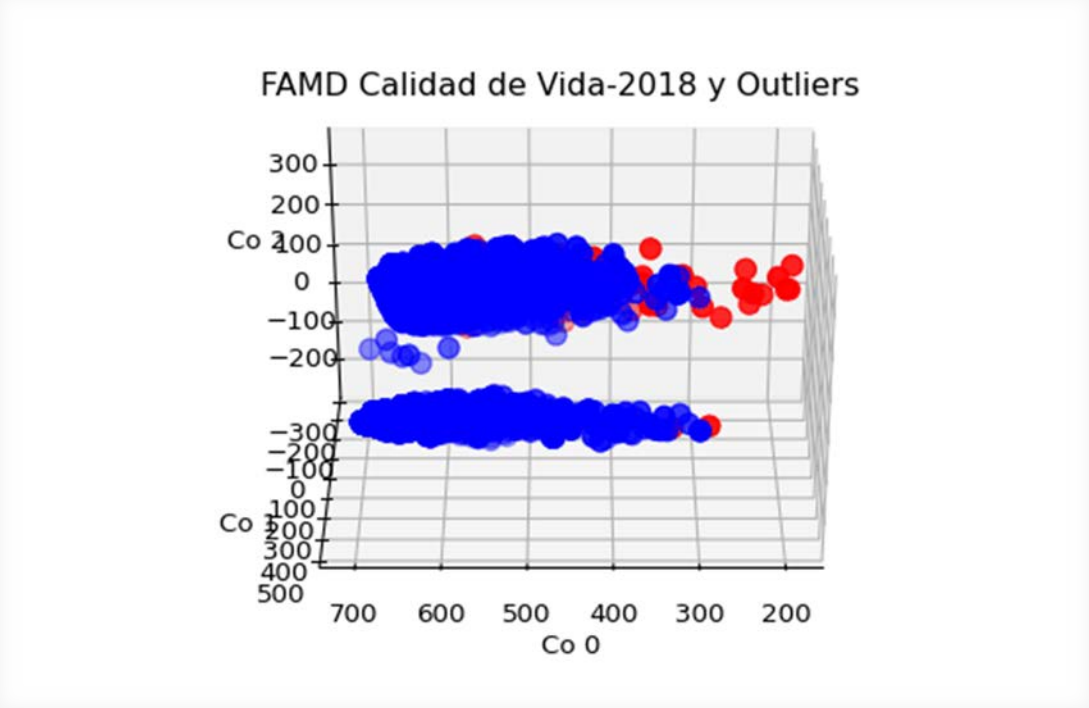
Identificación de Outliers utilizando distancia de Gower



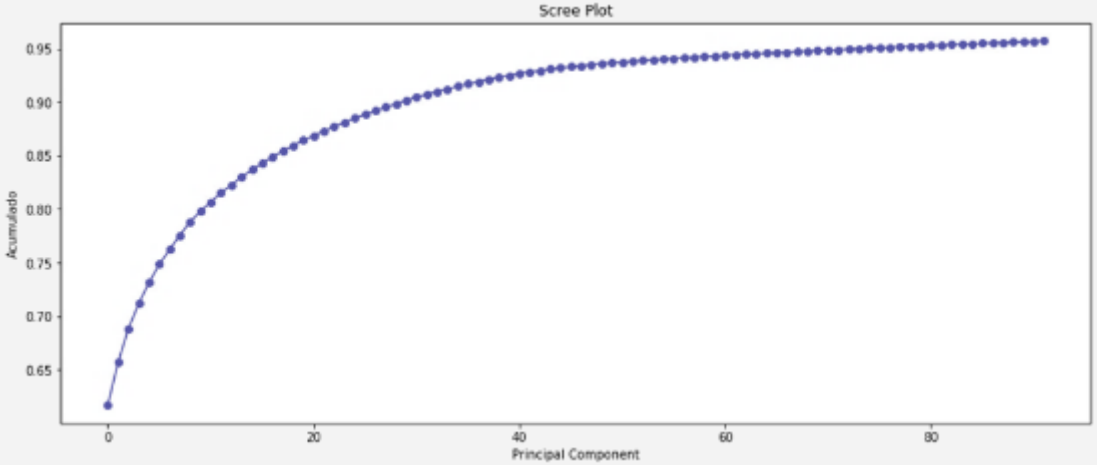


Se obtuvieron 153 hogares como outliers, lo que representa un 1.69% de *outliers* en el total de datos.

Gráfica de FAMD con los outliers



Reducción de dimensionalidad
FAMD



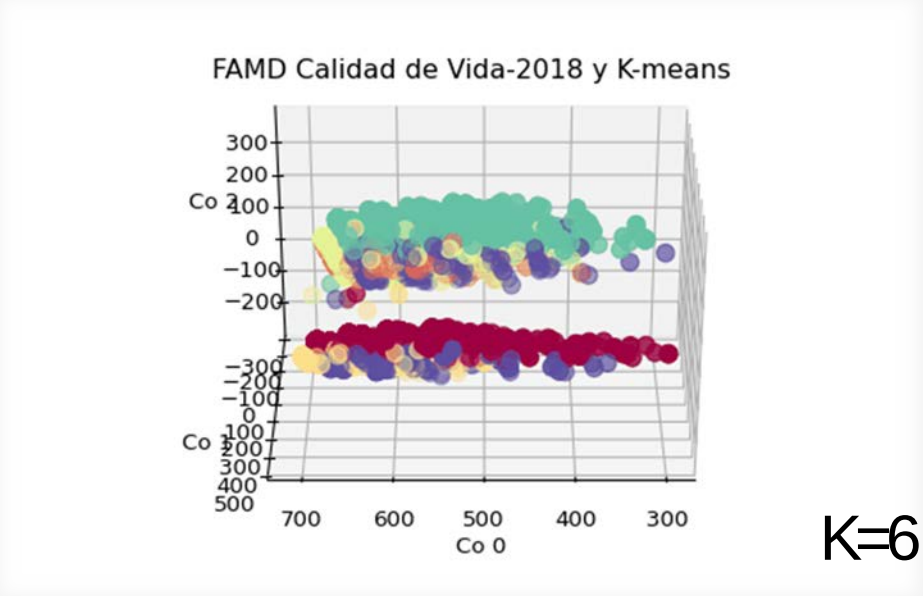
17

Componentes

85%

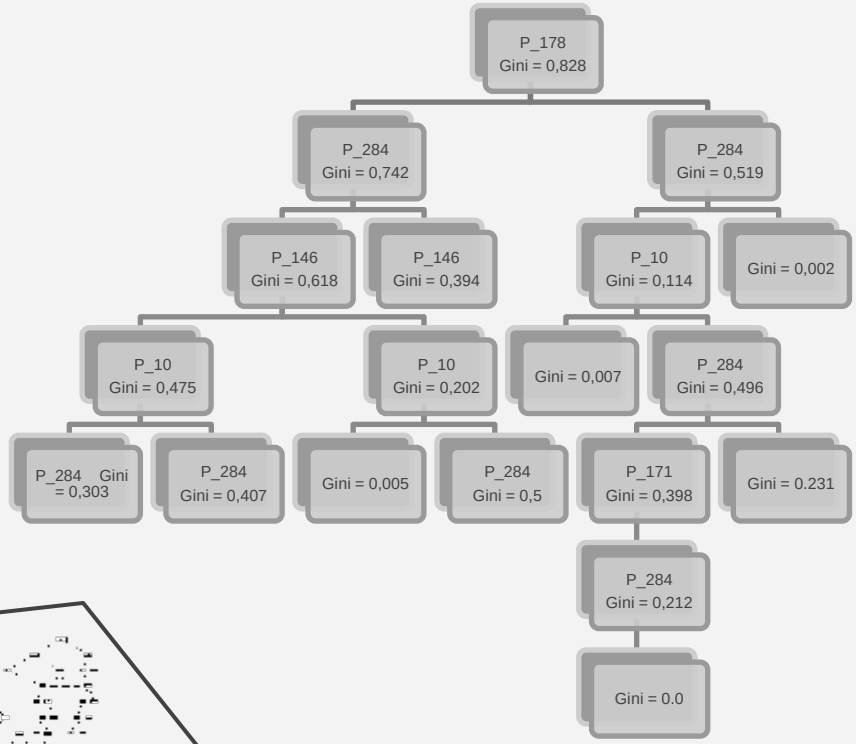
Variabilidad

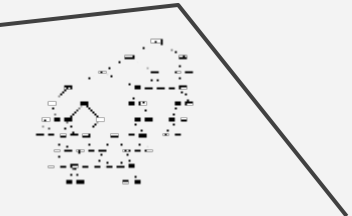
K-Means FAMD



K=6

Modelo árboles de decisión





Puntajes de precisión

Clasificación	Data	Score testeo	Score entrenamiento
K-Means	DF Original	0,98	0,99

Importancia de las variables en el modelo

Valores	Labels
0,492258	p_284
0,23661	p_146
0,187391	p_178
0,072025	p_10
0,003672	p_285

Valores promedio

Kmean_ labels	Gasto per cápita	Estudios universitarios	Posgrado	Vehículos	Electrodomésticos	Personas
4	$590.000	0,7	0,3	0,5	13,0	3,0
3	$367.083	0,4	0,1	0,3	12,0	3,0
2	$336.667	0,3	0,1	0,2	11,0	3,0
0	$297.750	0,3	0,1	0,2	12,0	4,0
1	$230.000	0,1	0,0	0,0	7,0	3,0
5	$225.000	0,1	0,0	0,0	7,0	3,0
Total	$314.750	0,3	0,1	0,2	10,0	3,0

Al analizar los resultados de las etiquetas de K Means con los valores medios de algunas de las variables socioeconómicas, se observan ciertos rasgos característicos para cada etiqueta que permiten generar diferenciación entre estas, y percibir progresividad en la mayoría de variables, como es el caso del gasto per cápita.

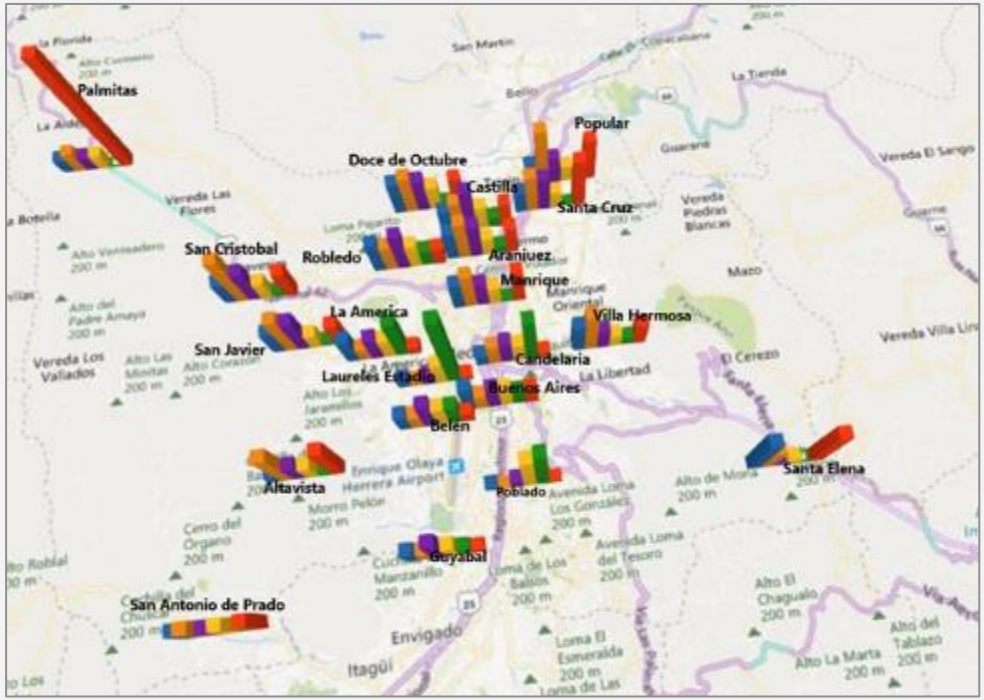
Distribución de los hogares de cada estrato por nivel de Kmeans

Kmean_ labels	1	2	3	4	5	6	Total general
4	1%	2%	10%	34%	56%	44%	14%
3	9%	12%	16%	21%	19%	40%	16%
2	14%	20%	25%	20%	8%	10%	19%
0	13%	20%	19%	13%	8%	4%	16%
1	35%	26%	17%	5%	4%	0%	19%
5	28%	21%	12%	6%	5%	2%	15%
Total	100%	100%	100%	100%	100%	100%	100%

Cuando se analiza la distribución de los hogares de cada estrato por nivel de K Means, se evidencia que en las etiquetas 4 y 3 tienden a concentrarse la mayoría de los hogares de los estratos más altos (5 y 6), representando el 76% y 84% de los hogares, respectivamente.

Cantidad de hogares de K-Means por comunas

- Grupo 0
- Grupo 1
- Grupo 2
- Grupo 3
- Grupo 4
- Grupo 5



De forma visual se puede apreciar cómo es la agrupación de los hogares realizada por K-means, ubicados por la respectiva comuna.
Se destaca como la mayoría de los hogares del corregimiento San Sebastián de Palmitas están en el grupo 5.

05 CONCLUSIONES

- Una clasificación socioeconómica debería considerar, de manera simultánea, variables del hogar, de los individuos que lo conforman y del entorno.

- Una de las etapas más complejas y crítica es la exploración y análisis de los datos. Es importante identificar claramente el tipo de datos disponible porque esto determina la estrategia para su análisis

- La distancia de Gower y la reducción de dimensionalidad FAMD son herramientas de gran utilidad para analizar datos mixtos. Estas herramientas fueron utilizadas para identificar outliers (distancia Gower) y reducción de dimensionalidad para realizar clasificación no supervisada (FAMD) lográndose resultados satisfactorios y claros.

- Es de destacar que no existen en la actualidad amplia documentación de artículos y librerías en Python de modelos que incluyan variables categóricas y numéricas, convirtiéndose en una oportunidad para generar nuevos algoritmos que permitan mejores desempeños.

GRACIAS