PROYECTO INTEGRADOR

Clasificación socioeconómica de hogares de Medellín

Presentado por:
Susana Londoño Muñoz
David Betancur Londoño
Catalina Piedrahita Jaramillo
Diego Andrés Jaramillo Zapata
Diego Andrés Valderrama Laverde

Profesores:
Henry Laniado Rodas
Edwin Nelson Montoya Múnera
José Antonio Solano Atehortúa
Juan Guillermo Lalinde Pulido

Universidad EAFIT
Maestría en Ciencias de los Datos y Analítica
Medellín
2020-Semestre I

Contenido

Resumen

A partir del conjunto de datos de la Encuesta de Calidad de Vida de Medellín 2018, y considerando variables de hogar, vivienda, individuo y entorno, se hizo una modelación que permitió clasificar los hogares por grupos similares a nivel socioeconómico.

El punto de partida fue la contextualización de algunas clasificaciones socioeconómicas existentes entre las cuales se destaca el estrato socioeconómico que clasifica la vivienda y no los individuos que la habitan. Así, se planteó el alcance del análisis y los pasos a ejecutar que se fueron afinando de forma iterativa.

Posteriormente, se ejecutó la adquisición y entendimiento de los datos utilizados. Se realizó la búsqueda de fuentes de información, definición de un esquema de trabajo y ciclo de vida de los datos. La naturaleza de los datos identificados conllevó el uso de la distancia de *Gower* para la eliminación de *outliers*, creación de variables ficticias (*dummies*) y reducción de dimensionalidad a partir de la implementación del método *Factor Analysis of Mixed Data* (*FAMD*, por sus siglas en inglés).

Una vez ajustados los datos, se aplicaron modelos no supervisados y supervisados. En el primer caso, se obtuvieron resultados para *K-Means* y *K-Medoids*, asignando para cada modelo seis niveles de clasificación. Para *K-Means* se partió del resultado de la reducción de dimensiones del método *FAMD* mientras que en *K-Medoids* se utilizó el *dataframe* original con la creación de variables ficticias (*dummies*). Finalmente, con los modelos supervisados de *Árboles de decisión*, *Regresión Logística* y *Naive Bayes*, se evaluó la consistencia de los modelos no supervisados, sobresaliendo la alta precisión que se obtuvo con *Árboles de decisión* y *Naive Bayes* para el resultado de *K-Means*.

Palabras claves: clasificación socioeconómica; encuesta de calidad de vida de Medellín; distancia *Gower*; *FAMD*; modelos no supervisados; modelos supervisados.

Descripción del problema

En términos generales, identificar grupos poblacionales con características similares puede llegar a representar, tanto para entes públicos como privados, un medio o instrumento para alcanzar una amplia variedad de objetivos económicos y sociales, dentro de los cuales sobresalen:

- Determinar nichos de mercado para su aprovechamiento comercial.
- Focalizar subsidios.
- Establecer esquemas tarifarios de productos y servicios.
- Priorizar y desarrollar proyectos de inversión.

- Diagnosticar territorios.

Para esto, se han utilizado todo tipo de métodos estadísticos y analíticos, a partir de diferentes variables, no solo de caracterización de individuos, sino de sus entornos, incluyendo temas de hogar, vivienda e inclusive de hábitat.

En el caso colombiano, una de las clasificaciones más utilizada es la relacionada con el estrato socioeconómico de la vivienda, cuya finalidad ha sido la de establecer un esquema tarifario diferenciado para el pago de servicios públicos domiciliarios. Se han definido 6 estratos socioeconómicos, donde a medida que se avanza en la escala ascendente se tienen cargos superiores, bajo la premisa de que quienes tienen mayor capacidad de pago deberían pagar servicios públicos más altos y contribuir a que los estratos bajos puedan pagar sus facturas. Los estratos 1, 2 y 3 corresponden teóricamente a estratos bajos que albergan a los usuarios con menores recursos, los cuales son beneficiarios de subsidios en los servicios públicos domiciliarios; los estratos 5 y 6 corresponden a estratos altos que albergan a los usuarios con mayores recursos económicos, los cuales deben pagar sobrecostos (contribución) sobre el valor de los servicios públicos domiciliarios. Mientras que el estrato 4 paga exactamente el costo de prestación del servicio.

Sin embargo, la estratificación no ha sido ajena a críticas, entre las que sobresalen, por ejemplo, que representa una metodología de segregación socioespacial, donde se acentúan las diferencias territoriales al interior de los municipios, polarizando los grupos sociales. Por otro lado, algunos estudios han demostrado que en algunos casos no hay correlación positiva entre capacidad de pago y estrato, lo que hace que la focalización de subsidios sea ineficiente, dado que existirán hogares que se beneficiarán del subsidio pese a tener capacidad de pago. Todo esto, porque, metodológicamente, la estratificación no considera las suficientes variables de caracterización socioeconómica y porque recae sobre el inmueble, y no los hogares, que son los agentes activos[1].

Destacando, además que, en algunos casos, la estratificación ha sido utilizada para definir la pertenencia de los hogares a clases sociales: baja, vulnerable, media y alta. Inclusive, ha llegado a ser utilizada para el cobro de tarifas diferenciadas en cuanto al pago de costos educativos, el acceso a subsidios estatales de todo tipo, que incluyen el acceso a espacios recreativos, campañas comerciales o el cobro de valorización.

Asimismo, en Colombia, se han utilizado otras métricas de clasificación de los hogares, como, por ejemplo, la de clase social, la cual se basa en la metodología del Banco Mundial

[1] Por ejemplo, esto se puede evidenciar en los siguientes estudios:
Departamento Nacional de Planeación (2008). Evaluación de la estratificación socioeconómica como instrumento de clasificación de los usuarios y herramienta de asignación de subsidios y contribuciones a los servicios públicos domiciliarios.
Betancur, D. (2012). Modelo de capacidad de pago para categorizar usuarios de servicios públicos de agua y saneamiento básico.
Alcaldía de Bogotá (2016). La Estratificación en Bogotá. Impacto social y alternativas para asignar subsidios.

y que se sustenta en la clasificación de los hogares según su ingreso per cápita. Con esta, se define si un hogar es de clase baja, vulnerable, media o alta. Sin embargo, se han hecho a nivel de muestras estadísticas y no de censo, dada la dificultad de obtener la variable de ingresos, tanto por la renuencia del informante como por la dificultad de acceder a otras fuentes.

Bajo esta misma línea, está la clasificación por categoría tarifaria, que está circunscrita al sistema de compensación familiar, donde los trabajadores y sus beneficiarios se clasifican como TA (si tienen salarios mensuales entre 0 y 2 SMMLV), TB (salarios mensuales entre 2 y 4 SMMLV), y TC (más de cuatro SMMLV). Aunque si bien se cuenta con la variable objetivo, deja de lado otras variables de caracterización, además, de que, en el caso colombiano, los niveles de formalidad son cercanos al 50%, por lo que no se lograría cubrir al total de población. Además, de dificultades para su georreferenciación.

Por otro lado, con el fin de identificar población pobre o en alto grado de vulnerabilidad, desde el Gobierno Nacional se han creado diferentes instrumentos, como es el caso del Sisben, partiendo de información primaria, e incluyendo otro tipo de variables de caracterización, pero dejando de lado, a un porcentaje significativo de la población.

Bajo este contexto, es que se propone para los hogares de Medellín, probar la modelación de variables socioeconómicas, que incluye caracterización de los individuos que lo conforman, vivienda que cohabitan y variables de entorno, para determinar grupos sociales y hacer clasificaciones que posibiliten encontrar nichos de mercado, focalizar intervenciones y encontrar similitudes en términos de los hogares y personas que los habitan.

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.

Adquisición y entendimiento de los datos

Fuentes de datos

La búsqueda inicial de información incluyó la exploración de fuentes de datos abiertos con información socioeconómica de los hogares de Medellín y municipios del área metropolitana. Algunos de los conjuntos de datos que se consideraron inicialmente incluyen información del mercado laboral, datos inmobiliarios, zonas de riesgo, zonas geoeconómicas, atractivos turísticos, espacio público, sisben y calidad de vida del Área Metropolitana y Medellín. De esta exploración, se identificó que el municipio que provee más información abierta es Medellín y que de los conjuntos de datos el más completo era el de la medición de Calidad de Vida.

Se utilizó el conjunto de datos de la **Encuesta de Calidad de Vida de Medellín (ECV).** El objetivo de esta encuesta es obtener información confiable y oportuna, expresada estadísticamente sobre variables físico-espaciales, sociales y demográficas, referida a cada una de las comunas y corregimientos de Medellín. El marco muestral fueron viviendas por estrato socioeconómico, comuna y corregimiento registradas en la base de datos de puntos de servicio suministradas por EPM. El tamaño muestral para 2018 fue: 9196 viviendas, 9228 hogares y 30941 personas en comunas y corregimientos. Además, la encuesta indaga por 342 variables, de vivienda, hogar y personas.

Esta información es de acceso público, disponible digitalmente en la siguiente página https://bit.ly/2ZfcCrA, cuyo dominio es de la Alcaldía de Medellín.

Ambiente tecnológico

Para el desarrollo del proyecto, se utilizó el sistema de archivos compartidos Google Drive, donde se desarrolló una estructura de carpetas colaborativa. Se basó en la metodología propuesta por Microsoft, Team Data Science Process (TDSP), a continuación, se muestra la estructura de carpetas:

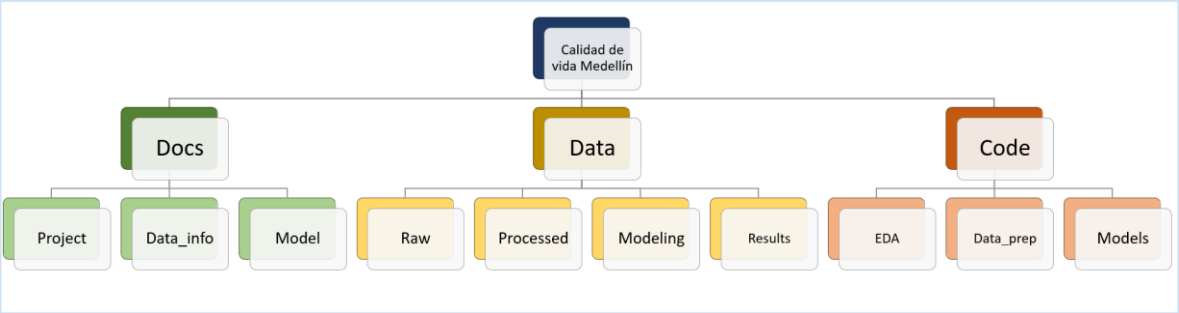

Figura 1.

En el directorio ***Docs***, subdividido en: *project, data_info* y *model*, se encuentra la información de gestión del proyecto (agenda, delegación de tareas y documentación), los metadatos y la bibliografía de los modelos. Los documentos presentados son archivos de texto. Los relacionados a la gestión del proyecto fueron realizados por los miembros del proyecto de manera colaborativa en los formatos disponibles en Google Drive. Los metadatos se obtuvieron de la fuente de información y de algunos *scripts* de análisis exploratorio de los datos.

En ***Data*** (ver figura 2), se encuentran almacenados todos los *datasets* en diferentes estados de procesamiento. El formato usado es CSV, utilizando como separador el carácter ";", ya que algunos valores dentro del conjunto de datos podrían contener ",".

Los datos inicialmente se almacenaron en *raw*, al realizar exploración y transformaciones iniciales, pasaron a *processed*, cuando se retiran los *outliers* y se realizan transformaciones para poderlos entrar a los modelos, se guardaron en *modeling* y luego los resultados de los modelos iniciales que incluye conjuntos de datos que pueden ingresar a otros modelos, se almacenaron en *results*.

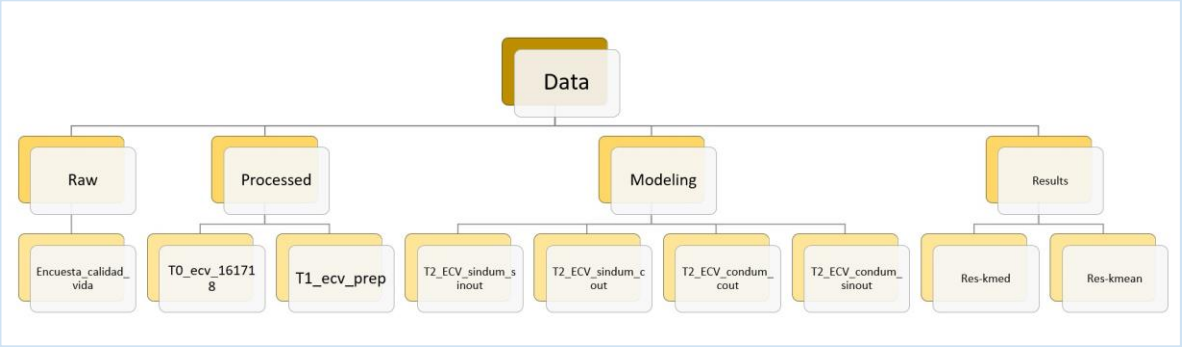

Figura 2.

Finalmente, en ***Code*** (ver figura 3), se encuentran tres carpetas (*EDA, data_prep y models)* donde se almacenaron los *scripts* usados para la exploración, transformación y modelado de los datos. La entrada de los *scripts* son los datos almacenados en las diferentes subcarpetas de *Data* y las salidas también serán archivos de datos ya transformados que se almacenarán en otra subcarpeta (ver figura 4).

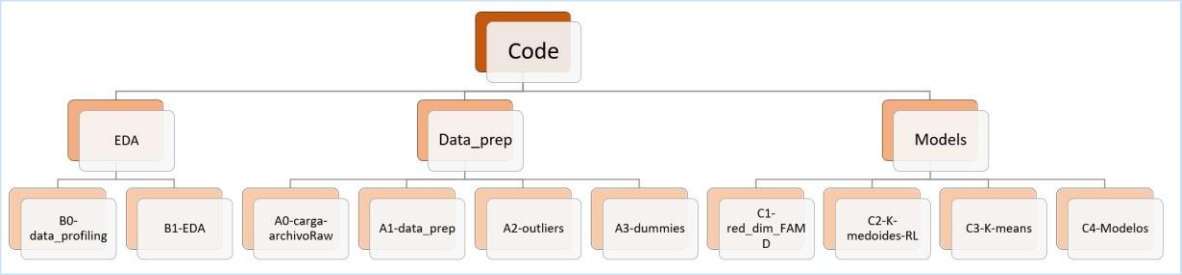

Figura 3.

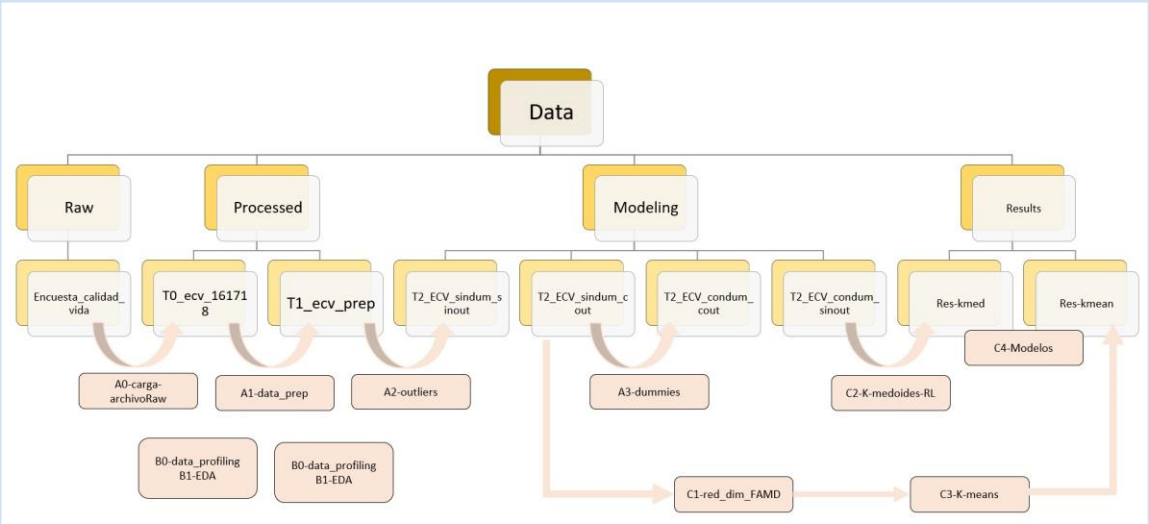

Figura 4.

El lenguaje usado para el desarrollo del proyecto fue Python. Se utilizó la herramienta Google Colab que permite escribir y ejecutar código de Python en el navegador, teniendo la facilidad de interactuar con los datos almacenados en Google Drive y la ejecución con capacidad computacional en la nube.

Los *notebooks*, es decir, los archivos de código en Colab, tienen el formato: *ipynb*, por lo tanto, son archivos que no solo contienen código, sino también la posibilidad de agregar texto e imágenes, facilitando la documentación del código.

Se utilizaron diversas librerías de Python, entre ellas *pandas, numpy, pyspark, sci-kit-learn, scipy, statsmodels, matplotlib, seaborn, gower, prince, imageio, mlxtend* y *collections*, de estas cabe resaltar *gower* que permitió desarrollar una distancia apropiada para el tipo de datos que se tenían, categóricos y numéricos, y *prince* que permite la aplicación de un método de reducción de dimensionalidad para los datos mixtos.

Además, para el despliegue final del proyecto se puso de manera pública la estructura de carpetas mencionada en el *github* del equipo: (https://github.com/SusanaLondono/PI-CDS-2020-1). Mediante la extensión *Git Large File Storage (git lfs)*, que permite almacenar archivos grandes en los repositorios, se pudieron almacenar los *datasets*.

Ciclo de vida de los datos

Tras obtener los datos, se debe realizar un proceso de entendimiento, preparación y limpieza para obtener *datasets* que cumplan los requisitos para entrar a los modelos, no solo deben ser datos de cierto tipo, sino que deben ser datos "limpios", para tal objetivo, se fue desarrollando un flujo de trabajo teniendo como base la estructura de carpetas y el ciclo general de un proyecto de ciencia de datos ver figura 5, como lo indica la metodología TDSP.

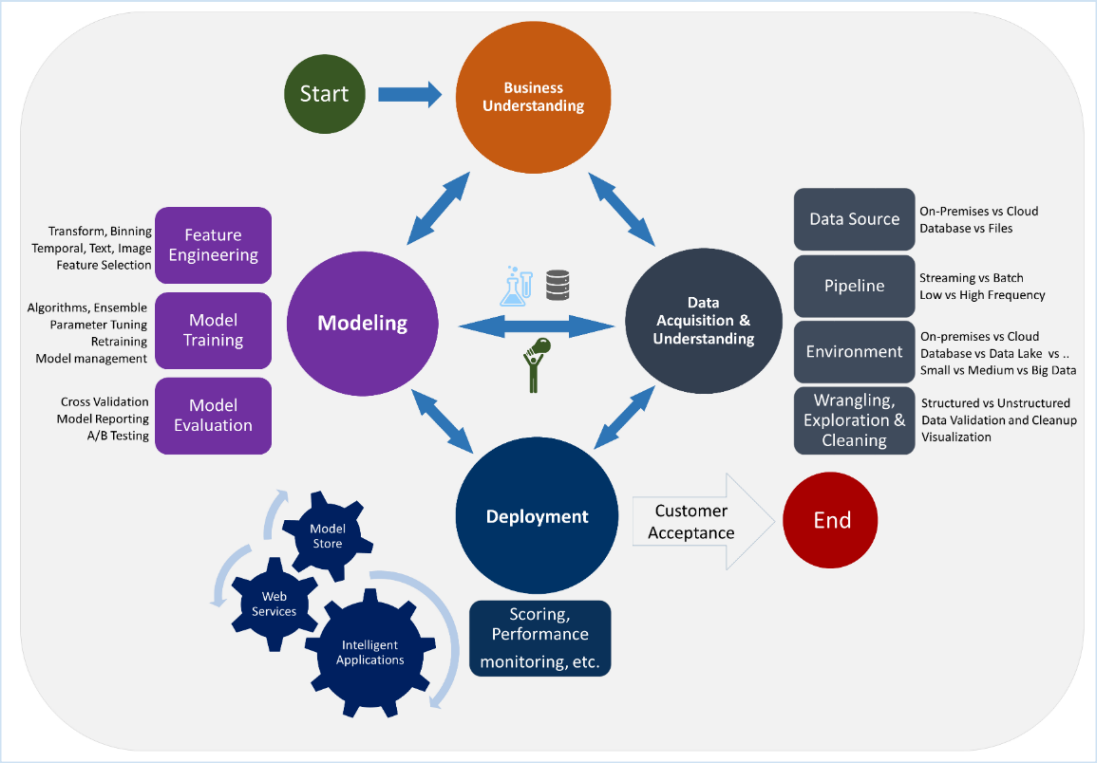

Figura 5.

En el subdirectorio ***raw*** se encuentran los datos crudos (encuesta_calida_vida.csv), es decir, tal y como se obtienen de la fuente. Estos, deben pasar por procesos de exploración y transformación (A0-carga_archivo, A1-data_preparation, B0-data_profiling y B1-EDA), de allí se generaron nuevos archivos de datos, analizados posteriormente, con algunas transformaciones necesarias (T0-ECV_161718.csv y T1-ECV-prep.csv) y finalmente almacenados en ***processed***.

Antes de que los datos estuvieran listos para ingresar a los modelos, primero se debió realizar el análisis de valores atípicos (*outliers*), apropiado al tipo de datos que se está trabajando (A2-*outliers*), y retirar dichos valores, generando conjunto de datos, para este caso, dos archivos ordenados según la distancia al vector centro, uno con valores atípicos y otro sin estos (T2-sindum-conout.csv y T2-sindum-sinout.csv), de este modo pasaron a almacenarse en ***modeling*** donde pudieron ser tomados para modelar.

Para unos modelos fue necesario otro proceso debido a la naturaleza de los datos analizados. Se crearon variables ficticias, esto aplica para las variables categóricas presentes (A3-dumimes_categoricas) y se obtuvieron otros dos *datasets* (T2-condum-conout.csv y T2-condum-sinout.csv) que quedaron almacenados en el directorio ***modeling***.

Con los *datasets* disponibles en la carpeta ***modeling***, se realizaron los primeros modelos (C1-K-medoides-RL y C2-K-means), los cuales arrojan unos resultados (R2-kmedoides, R2-*FAMD* y R3-kmeans) que luego se usaron en los demás modelos, estos resultados se depositaron en la carpeta ***results***. Ellas figuras 6 y 7 se resume el flujo de los datos:

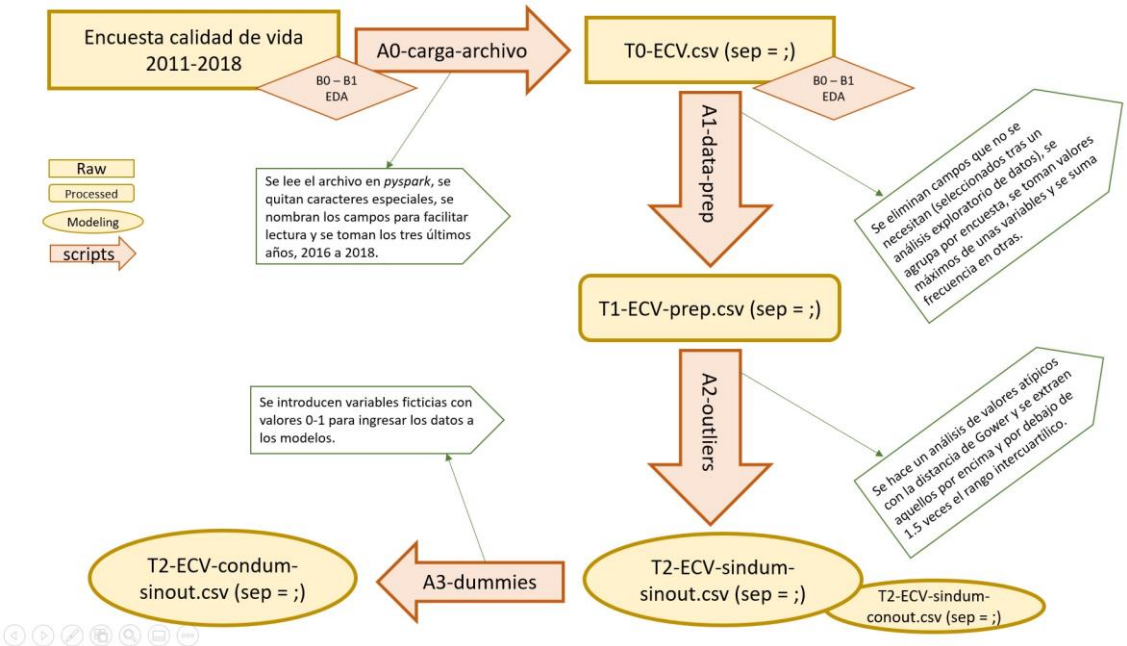

Figura 6.

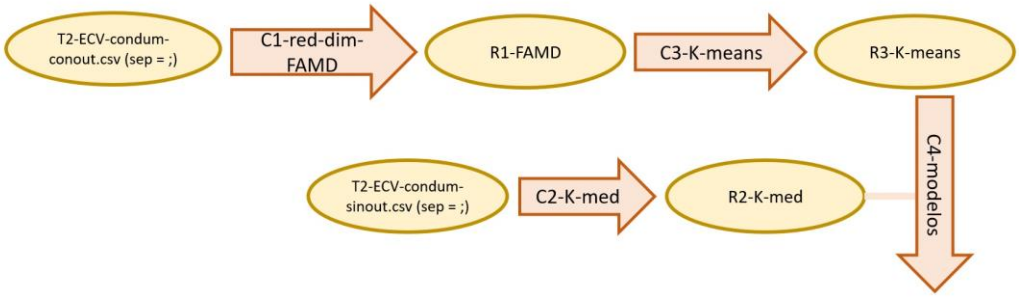

Figura 7.

Entendimiento de los datos

El entendimiento de los datos es un paso crítico previo a la preparación para un posterior análisis. Esta tarea requiere un porcentaje importante del tiempo y esfuerzo total.

El entendimiento de la estructura y contenido se inició con la revisión de los conjuntos de datos, diccionarios de datos y otros archivos de metadatos obtenidos de la fuente.

Posteriormente, se hizo perfilamiento de los campos de los conjuntos de datos y se obtuvo información general y detallada. A modo de ejemplo, de los análisis ejecutados se muestran algunos resultados del perfilamiento del archivo: *T0_ECV_18.csv*

Número de filas	30.934	**Tipo float64**	14
Núm. de columnas	369	**Tipo int64**	353
Tipo object	3	**Nulos**	0

Tabla 1.

En el análisis detallado de cada campo y dependiendo del tipo de dato, se obtuvieron tablas resumen y gráficas. El campo estrato tiene el siguiente análisis:

Campo: Estrato	**Tipo dato** : int64	**Contador valores** : 30,934
Valores únicos : 6 (0.02%)	**Non-Null**: 30,934 (100%)	**Valores Null** : 0 (0.00%)
Mín: 1.00	**Media**: 2.70	**Máx :** 6.00

Percentil 25 : 2.00	**Percentil 50** : 2.00	**Percentil 75**: 3.00

Tabla 2.

Análisis gráfico:

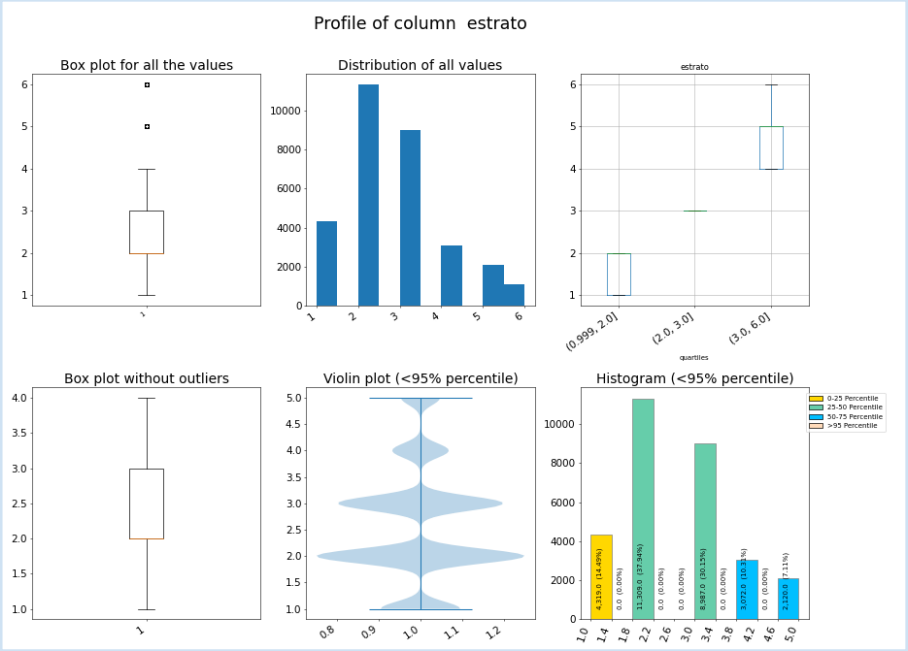

Figura 8. Este primer análisis se generó de forma automática. Posteriormente esta variable se clasificó como categórica.

El resultado detallado del perfilamiento de los campos de los conjuntos de datos se encuentra en los documentos: *perfilado_T0.docx* y *perfilado_T2_ECV_sindum_sinout_ord.docx*

Finalmente, considerando la importancia de los datos categóricos para el análisis, se construyó un listado para indicar si las preguntas contenían datos categóricos y un filtro de las preguntas a excluir del análisis. Los archivos con el diccionario y otros archivos de metadatos se encuentran en las carpetas */datasets/metadata*

Preparación de datos

Outliers

En la preparación de los datos es esencial realizar una identificación de los *outliers* porque no es ideal incluir este tipo de datos cuando se van a construir los modelos. Dado que el

conjunto de datos contaba con variables mixtas, para realizar la identificación de los *outliers* se utilizó la distancia de **Gower**, que permite evaluar variables cualitativas y cuantitativas calculando la distancia manhattan para las numéricas y dice para las categóricas.

A continuación, la fórmula matemática que usa la función de Python para calcular la distancia de *Gower*:

$$S_{ij} = \frac{\sum_{k}^{n} w_{ijk} S_{ijk}}{\sum_{k}^{n} w_{ijk}}$$

- where:

S_{ijk} denotes the contribution provided by the $k-$th variable, and

w_{ijk} is usually 1 or 0 depending if the comparison is valid for the $k-$th variable.

Figura 9. Tomada de: https://bit.ly/3fYUkBH

Procedimiento realizado:

1. Se estandarizaron los datos cuantitativos de todo el *dataset.*
2. Se creó un vector centro que se usa como referencia base para calcular las distancias de todos los hogares a este, para construirlo se calculó la moda para las variables categóricas, y con los datos cuantitativos estandarizados se calculó la mediana.
3. Se calculó la **distancia de Gower** de todos los hogares a este vector centro. A continuación, se muestra el histograma y boxplot creado a partir de dichas distancias para el caso de los hogares en la ECV 2018.

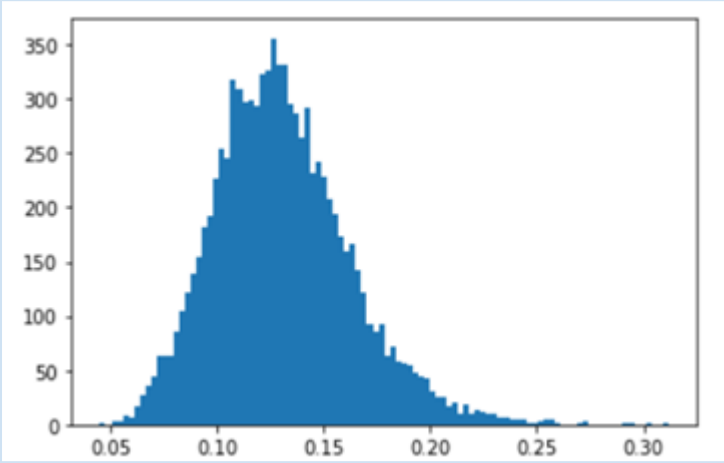

Figura 10. Histograma distancia de *Gower*: todos los puntos al vector centro

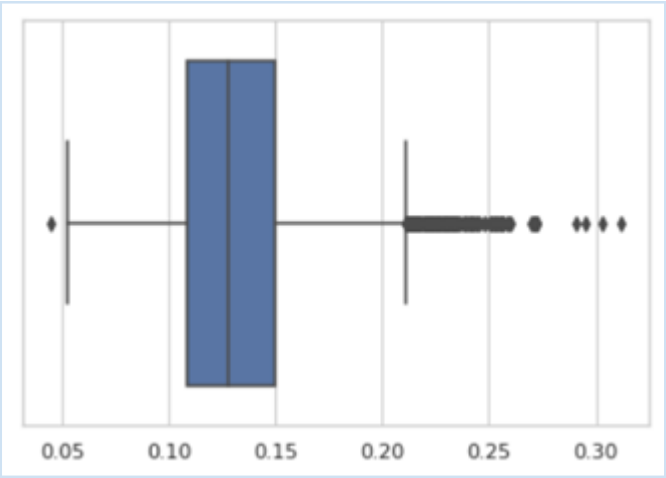

Figura 11. Boxplot distancia de *Gower*: todos los puntos al vector centro

4. Se analizaron las gráficas presentadas de las distancias para determinar los *outliers*. En el histograma se observó que estas distancias parecen tener una distribución normal. Usando el boxplot como base para determinar los *outliers*, se puedo ver en la gráfica algunos puntos por fuera de los bigotes (1.5 veces el rango intercuartílico), se identificó uno fuera del límite inferior y varios, la mayoría fuera del límite superior. Al tomar estos puntos como valores atípicos, se obtuvieron 153 hogares, lo que representa un 1.69% de *outliers* en el total de datos.

Con la tabla resumen de medidas de tendencia central de estas distancias, (ver figura 12), se pudo obtener el coeficiente de variación [(desviación típica/media) x100], el cual es 24.5%, demostrando que la media aritmética era representativa del conjunto de datos. Tal información permitió intuir que el vector centro calculado a partir de la mediana y la moda de los datos fue una buena aproximación al centro de estos porque la distancia de los hogares con respecto a él no tenía mucha variabilidad.

mean	std	min	25%	50%	75%	max
0.130729	0.031995	0.04503	0.108483	0.127906	0.149568	0.311587

Tabla 3. Estadísticos resumen distancias de *Gower* de todos los puntos al vector centro

Creación de variables ficticias (Dummies)

Para obtener datos que pudieran entrar a algunos modelos se tuvo que introducir un bloque de variables ficticias, este proceso se realizó para las variables categóricas presentes en el

conjunto de datos, cada categoría se puso como variable y los registros tomaron valores de 0 donde no se presentaba dicha categoría y 1 donde se presentaba. A continuación, una tabla de ejemplificación de la transformación.

Sin dummies	
hogar	p_10
0	1
1	2
2	2
3	3
4	4
5	6
6	5
7	4
8	3
9	2

Con dummies						
hogar	p_10_1	p_10_2	p_10_3	p_10_4	p_10_5	p_10_6
0	1	0	0	0	0	0
1	0	1	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	0	1
6	0	0	0	0	1	0
7	0	0	0	1	0	0
8	0	0	1	0	0	0
9	0	1	0	0	0	0

Tabla 4.

Este proceso se llevó a cabo en el *script* A3-*dummies*_categoricas, cuya salida son los *datasets* T2-condum-conout.csv y T2-condum-sinout.csv, el último fue el ingreso a K-medoides.

Reducción de dimensionalidad (*FAMD*)

Debido a la particularidad de los datos mixtos, no fue posible utilizar métodos tradicionales como análisis de componentes principales (PCA, por sus siglas en inglés). Por lo tanto, se utilizó el Análisis Factorial para Datos Mixtos (*FAMD*, por sus siglas en inglés), que permite analizar la similitud de observaciones a pesar de la presencia de diferentes tipos de variables. Es una combinación entre el análisis de componentes principales y el análisis de correspondencia múltiple. Esta metodología también está basada en la descomposición de valores singulares. La función de Python que se utilizó fue *prince.FAMD*, la cual utiliza randomized SVD de sklearn. Esta versión aleatoria de SVD es un método iterativo controlado para mantener la eficiencia computacional. Para el ejercicio se hicieron 100 iteraciones porque el algoritmo convergía muy rápido. La descomposición en valores singulares permitió obtener la estimación de la variabilidad explicada por cada componente y el peso de cada una de las variables en las componentes.

Se redujo la dimensionalidad a 17 componentes, las cuales en conjunto explican el 85% de la variabilidad total de los datos. Además, permitió contar con variables numéricas para la implementación de los modelos posteriores.

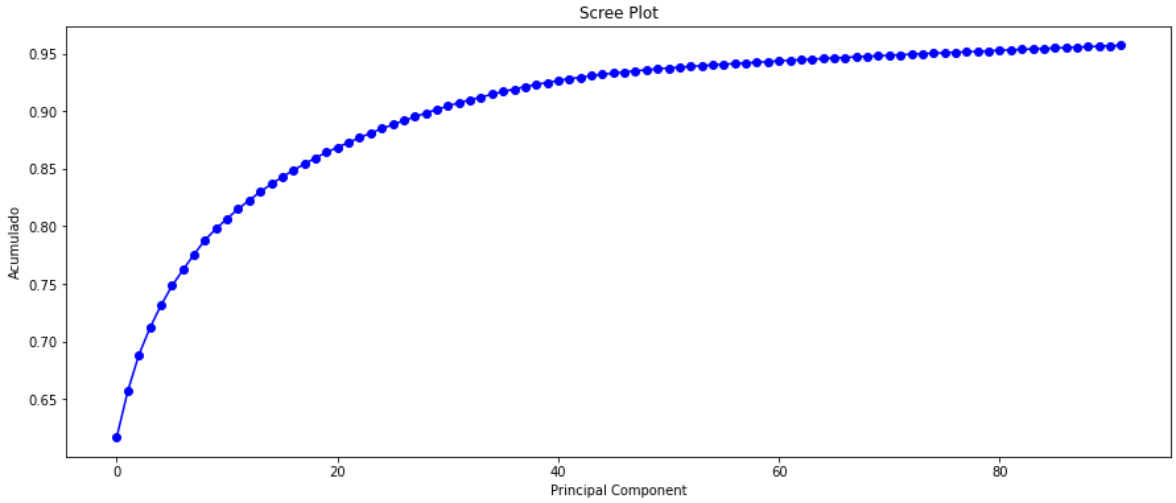

Figura 12. Acumulado de la variabilidad aportada por cada componente con *FAMD*

Modelos

A continuación, se presenta un esquema de la modelación realizada con la información procesada y analizada.

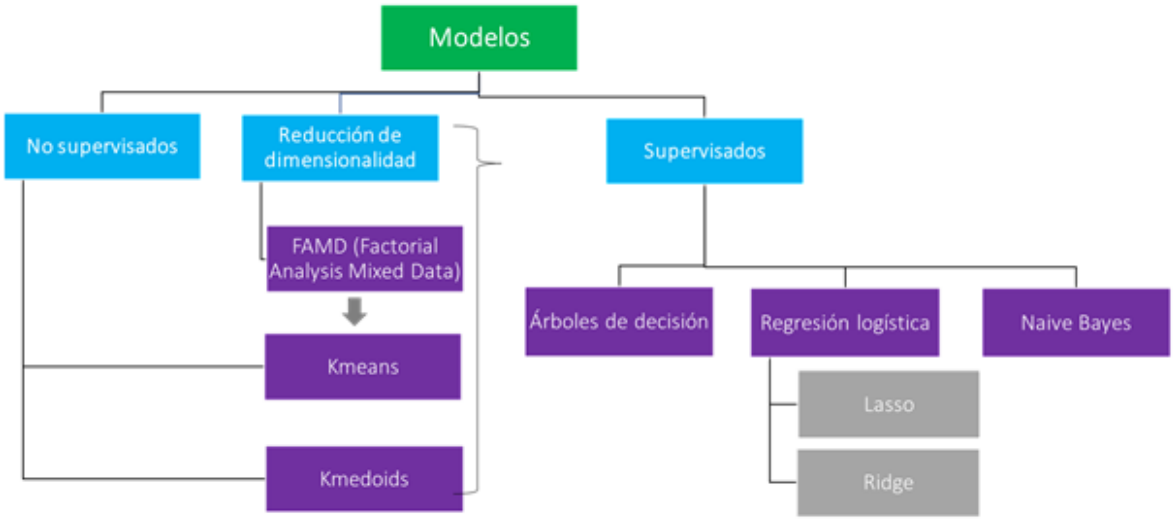

Figura 13.

Se parte de una premisa fundamental de que el entendimiento de un hogar a nivel socioeconómico es un tema multivariado, que involucra variables del hogar, de los individuos que lo conforman y del hábitat donde estos se desarrollan y materializan sus propósitos de vida.

Es así, como una vez finalizada la preparación de los datos, se procedió a analizar los modelos a implementar teniendo en cuenta la particularidad de los mismo, más específicamente haciendo referencia a la tipificación de las variables (numéricas y categóricas). Dentro de las investigaciones y el conocimiento obtenido en las materias vistas en el semestre, **se exploraron modelos supervisados y no supervisados**.

Se inició por la implementación de un algoritmo no supervisado llamado ***K-Medoides***, el cual fue implementado debido a la necesidad de calcular diferentes distancias dependiendo de la naturaleza de las variables; por consiguiente, se utilizó la **distancia de Gower**. Una de las razones para seleccionar ***K-Medoides*** es que el algoritmo utiliza puntos de los datos originales como centros y desde allí comienza la agrupación (clusterización) y el afinamiento de los centroides para obtener sus resultados. Se optó por identificar seis agrupamientos (K = 6) teniendo como base el mismo número de estratos socioeconómicos realizado por el municipio de Medellín.

Con los datos obtenidos de la reducción de dimensionalidad, las 17 componentes más significativas resultantes del ***FAMD***, se empleó otro método no supervisado, ***K-Means*.** También se definieron seis clusters (K=6).

Ahora bien, una vez se tienen dos hipótesis (***K-Medoides*** y ***K-Means***) de cómo se clasificarían socio - económicamente los hogares, evaluados en la encuesta de calidad de vida. Se procedió a validar la consistencia de estos resultados empleando modelos supervisados como los **árboles de decisión**, la **regresión logística** y ***Naive Bayes***.

Para iniciar el análisis, la siguiente tabla detalla los resultados de los diferentes escenarios evaluados con el método de **árboles de decisión**.

	Clasificación	Modelo	Data	Score testeo	Score entrenamiento
1	Kmeans	Árboles de decisión	DF Original - 92 variables	0.984513	0.993805
2	Kmeans	Árboles de decisión	FAMD - 17 componentes	0.971239	0.993510
3	Kmedoids	Árboles de decisión	DF Original - 488 variables	0.761504	0.843363
4	Kmedoids	Árboles de decisión	FAMD - 17 componentes	0.492478	0.616372

Tabla 5. Resultados árboles de decisión con diferentes fuentes

El primer escenario evaluado con árboles de decisión fue el conjunto de datos original con las 92 variables inicialmente identificadas como relevantes y la clasificación entregada por el método ***K-Means***. Los resultados fueron satisfactorios y se obtuvo una precisión del 99% con los datos de entrenamiento y un 98% con los de pruebas. Además, aprovechando la información adicional que entrega la ejecución de **árboles de decisión**, se identificaron

característica de mayor *significancia* para empezar a distinguir qué variables eran más relevantes. A continuación, el resultado para el caso 1.

	valores	labels
14	0.492258	p_284
4	0.236610	p_146
12	0.187391	p_178
1	0.072025	p_10
15	0.003672	p_285

Tabla 6.

Esto quiere decir que preguntas como la número *284 - ¿Cuales son los dos problemas más graves en orden de importancia para usted en relación con la seguridad que se presentan en su barrio?* tuvo una significancia del 49% para lograr la precisión del modelo. Mientras la pregunta *146 - Tipo de vivienda - Rancho, Cuarto, Inquilinato, Apartamento, Casa* tuvo una significancia del 23%.

De igual forma se aplicó el modelo de **regresión logística**, donde la variable dependiente fueron las agrupaciones (*clusters*) hallados tanto con ***K-Means*** como con ***K-Medoids***, y cuyas variables explicativas fueron las 17 componentes obtenidas de ***FAMD***. En el caso de la regresión logística con ***K-Means*** se obtuvo una precisión de 0,9988.

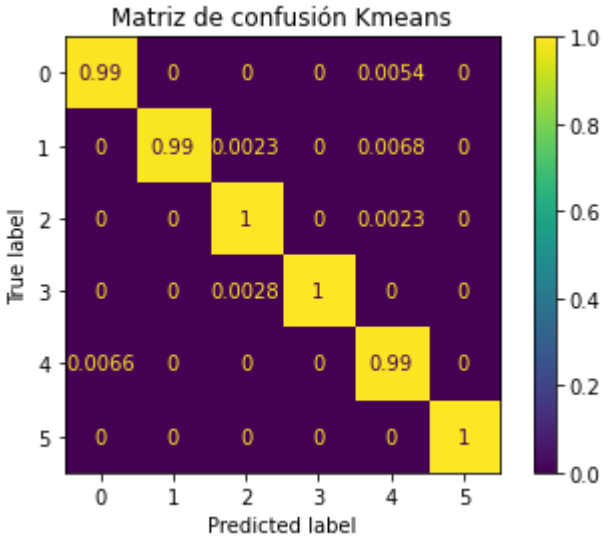

Figura 14.

Dada la presencia de *overfitting* se aplicaron los modelos de ***Lasso*** y ***Ridge***. Los cuales presentaron una reducción de la precisión, llegando ésta al 0,55, pero reduciendo considerablemente el número de variables explicativas.

En el caso de ***K-Medoids***, el modelo de regresión logística no presentó *overfitting*, aunque se obtuvo un nivel de precisión del 0,53, razón por la cual no se aplicaron ***Lasso*** y ***Ridge***.

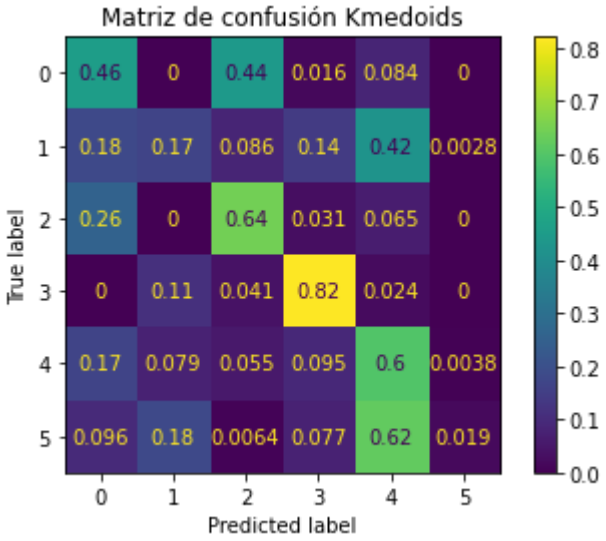

Figura 15.

En resumen, estos fueron los resultados tras la implementación de los modelos:

Clasificación	Modelo	Res_Error_Ent	Res_Error_Test	Número de Coef	Score
K-Means	Regresión logística	0.0	97.0	102	0.999
K-Means	Lasso	8637.3	2767.5	14	0.546
K-Means	Ridge	8493.6	2898.5	17	0.547
K-Medoides	Regresión logística	28388.7	6445.0	102	0.523

Tabla 7.

En el caso de ***Naive Bayes***, la precisión fue de 0.92 en pruebas y 0.93 en entrenamiento. Aquí se usaron, como entrada para entrenar el modelo, las 92 variables inicialmente identificadas como relevantes y la clasificación entregada por el método ***K-Means***.

	Clasificación	Modelo	Data	Score testeo	Score entrenamiento
2	Kmeans	Naive Bayes	DF Original - 92 variables	0.921239	0.936873

Tabla 8.

A continuación, los resultados de la matriz de confusión para el modelo.

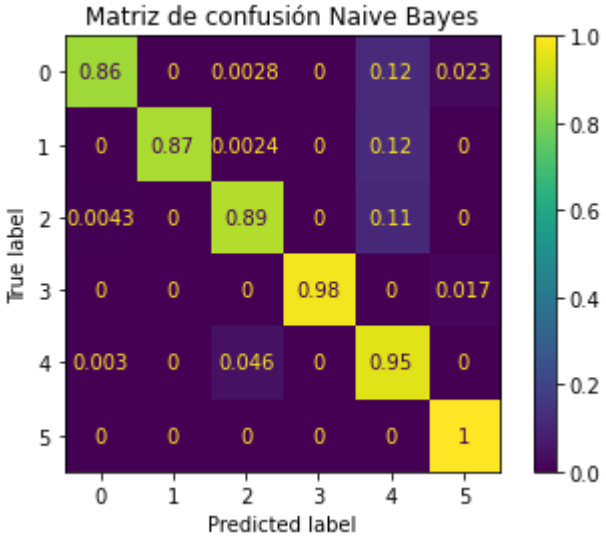

Figura 16.

Resultados

De los diversos modelos realizados, finalmente se toma como mejor clasificación de los hogares, la obtenida mediante *K-Means* usando las 17 primeras componentes, y como mejor modelo supervisado para estas etiquetas los árboles de decisión, dado que la precisión es alta y se puede apreciar la importancia de las variables para determinar la pertenencia a los grupos.

Es de resaltar que para entrenar los modelos supervisados se usó como variable respuesta las etiquetas del *K-Means* obtenidas mediante la reducción de dimensionalidad y como input todas las 92 preguntas identificadas como relevantes; la alta precisión al entrenar los modelos, puede dar evidencia de cómo a pesar de haber realizado una reducción de dimensionalidad para lograr la clasificación de los hogares, las 17 componentes describen muy bien las 92 variables. Al devolverse y usar todas las variables sin reducir dimensiones para obtener los grupos, se logran precisiones igualmente altas.

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, y que permiten generar diferenciación entre estas, y percibir progresividad en la mayoría de variables, como es el caso del gasto per cápita.

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

Tabla 9.

Cuando se compara 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.

	Estratos						
Kmean_labels	1	2	3	4	5	6	Total
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%

Tabla 10.

Recomendaciones y conclusiones

Una clasificación socioeconómica debería considerar, de manera simultánea, variables del hogar, de los individuos que lo conforman y del entorno. Considerar una sola dimensión puede limitar el alcance de su entendimiento e, inclusive, generar procesos de segregación socioespacial.

Uno de los principales retos de este proyecto consistió en trabajar con datos mixtos. Para este tipo de datos es importante validar cuando se trabaja en *python* que el tipo de variable esté correctamente definido porque esto tiene un impacto significativo en los resultados. En esta experiencia fue necesario definir un mecanismo para que las variables categóricas fueran reconocidas de esta manera.

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 aplicación de métodos no acordes a la naturaleza de los datos puede llevar a conclusiones erradas.

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 y alcances.

Referencias bibliográficas

- Alcaldía de Medellín (2020). Base de datos Encuesta Calidad de Vida 2018. Disponible en: https://bit.ly/2ZfcCrA.
- Alcaldía de Bogotá (2016). La Estratificación en Bogotá. Impacto social y alternativas para asignar subsidios. Disponible en: https://bit.ly/381192Q.
- Betancur, D. (2012). Modelo de capacidad de pago para categorizar usuarios de servicios públicos de agua y saneamiento básico. Disponible en: https://bit.ly/2ZeINaV.
- Chelaru-Centea, Nancy (2019). Calculate principal components of mixed-type data. Factor analysis of mixed categorical and continuous data in R and Python. Disponible en: https://bit.ly/2Yx9Sqz.
- Departamento Nacional de Planeación (2008). Evaluación de la estratificación socioeconómica como instrumento de clasificación de los usuarios y herramienta de asignación de subsidios y contribuciones a los servicios públicos domiciliarios. Disponible en: https://bit.ly/3g0mA6Y.
- Kesh, Sreemanto (2020) How to calculate *Gower*'s Distance using Python. Disponible en: https://bit.ly/3fYUkBH.
- Unrue, Matthew (2019). Analysis of Customer Churn in the Telco Customer Churn Dataset. Disponible en:https://bit.ly/3eAWNSJ.