Clustering de Clases Sociales en Colombia

Grupo 8 (MIAD – Aprendizaje No Supervisado)

Maria Camila Lenis, Oscar Álvarez, Ronaldo Ballesteros, Julio Solano

Resumen

Este trabajo explora la clasificación de las clases sociales en Colombia, utilizando técnicas de aprendizaje no supervisado como KNN, DBSCAN y clustering jerárquico. La clasificación oficial del DANE se basa en ingresos, pero ha sido objeto de debate, especialmente en cuanto a su precisión para reflejar la realidad social del país. El estudio emplea datos del DANE sobre pobreza y desigualdad, y, tras un proceso de limpieza y análisis, se identifican grupos basados principalmente en los ingresos, encontrando que otras variables como la edad o nivel educativo tienen poca relevancia en esta segmentación.

Los resultados muestran que, al aplicar clustering, la clase media definida por el DANE incluye grupos sociales muy diversos, desde vulnerables hasta clase alta. Este hallazgo sugiere que la clasificación actual no captura adecuadamente las diferencias socioeconómicas en Colombia, lo que podría impactar negativamente el diseño de políticas públicas. Se recomienda reconsiderar la definición de las clases bajas y medias para reflejar mejor las diferencias dentro de estas. Si bien la muestra utilizada es limitada, el estudio ofrece una nueva perspectiva sobre la estructura social del país y resalta la importancia de enfocarse en mejorar los ingresos de los hogares más vulnerables.

Introducción

La clasificación de las clases sociales es fundamental para el diseño de políticas públicas en Colombia. Actualmente, el DANE utiliza criterios de ingresos para definir estas categorías, pero los rangos salariales asignados han sido objeto de debate. Por ejemplo, se clasifica como clase media a quienes ganan entre $780.292 y $4′201.570 mensuales. Esto plantea la cuestión de si los ingresos son el único factor que define las clases sociales y si las categorías oficiales reflejan con precisión la realidad del país.

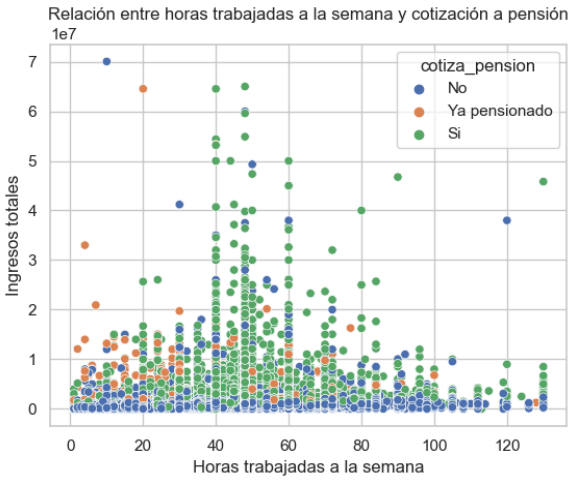
Este estudio busca aplicar algoritmos de aprendizaje no supervisado para identificar patrones en los datos socioeconómicos y comparar estos resultados con la clasificación del DANE. Al usar técnicas como el clustering, se espera ofrecer una nueva perspectiva sobre la estructura social en Colombia y contribuir a un diseño más adecuado de políticas públicas. Este enfoque ya ha demostrado ser útil en otros países, como Malasia, donde se ha empleado para identificar áreas de riesgo basadas en patrones socioeconómicos.

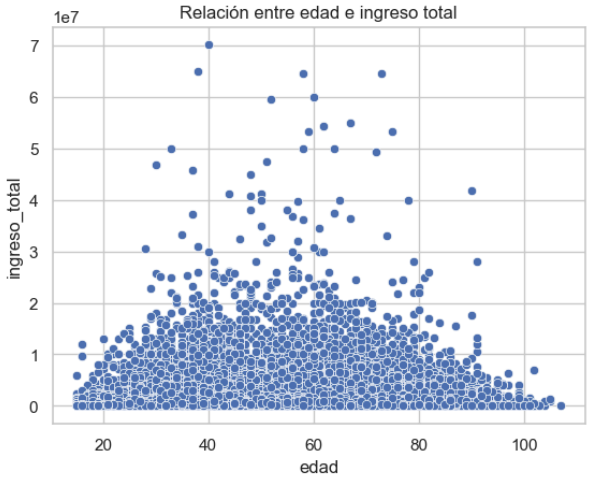
Una de las principales limitaciones para aplicar estos algoritmos es la desigualdad económica que prevalece en Colombia. La alta concentración de personas en las clases bajas y la presencia de valores atípicos podrían dificultar el análisis, ya que algoritmos como KNN pueden tener problemas para definir distancias entre individuos en términos de clase social. A pesar de estos retos, el estudio avanza considerando la complejidad del problema.

Materiales y métodos

Para el desarrollo del presente proyecto se utilizaron datos abiertos del DANE. En particular se utilizó el dataset de **Pobreza Monetaria y Desigualdad 2022**. Que contiene 132 variables relacionadas con el nivel de ingresos, información sociodemográfica, económica y ocupación para las personas. Se extrajo una muestra de este dataset, contando con *240.001* registros. De forma inicial se mapearon los nombres de las columnas codificadas siguiendo el diccionario de datos provisto por el DANE (DANE, 2022). Luego se decidió realizar un análisis exploratorio para comprender las relaciones en estos datos y así tomar decisiones sobre su tratamiento.

Se encuentra entonces que no existe una relación directa entre variables como edad, nivel educativo, horas a la semana trabajadas o si cotiza o no pensión, con el nivel de ingresos de una persona. De igual manera, se observa que hay una densidad grande de datos en los niveles más bajos de ingresos, acumulandose en rangos menores a $1’200.000, aún cuando se tiene datos atípicos que llegan a más de 7 millones de pesos.





Partiendo de las 132 variables se realizó el siguiente proceso de limpieza donde se eliminaron columnas con más de 200.000 datos nulos, se mapearon valores binarios a booleanos y se graficaron correlaciones entre distintas variables referentes a ingresos y nos quedamos solo con una puesto que se tenía una correlación de más del 97%. Esto quiere decir que teníamos datos

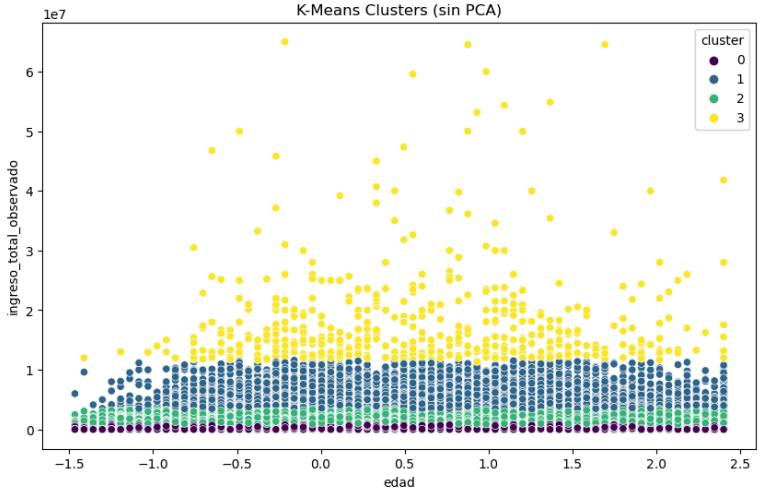
duplicados. Por otra parte, como la variable de interés es el ingreso, nos quedamos con 186,550 datos que correspondían al número de datos no nulos de esta variable.

Finalmente para pasar a la parte de los modelos de clustering se realizó codificación de variables categoricas y escalamiento de variables. Esto con el objetivo de que el algoritmo no fuera a darle más peso a algunas variables por encontrarse en escalas distintas, sobre todo si se trataba de valores como edad vs ingresos.

Con el objetivo de reducir la dimensionalidad y hacer un poco más sencilla la corrida de los algoritmos, se realizó PCA. Este método tuvo como resultados dos componentes principales que explicaban el 98% de la varianza. Al entrar a comprender a qué datos correspondía cada componente principal se contró que se trataban únicamente de las variables de ingresos e ingresos por arriendo. Lo que parece corresponder a lo encontrado en el análisis exploratorio. Realmente la edad, nivel educativo, pensión u ayudas, no hace la diferencia entre un individuo u otro. Lo hace el ingreso.

Resultados y discusión

Como se propuso inicialmente, se realizó un clustering utilizando KNN para formar cuatro clusters tal como lo realiza el DANE. Esto se hizo inicialmente sin reducir la dimensionalidad. Para conseguir visualizar las caracteristicas de cada cluster se realizó un diagrama de dispersión para la edad y el ingreso total. Como se puede apreciar en el gráfico, los clusters están formados únicamente teniendo en cuenta los ingresos, puesto que se distribuyen uniformemente por todas las edades como habíamos visto en el análisis exploratorio. Sin embargo, gracias a la presencia de outliers, se tiene un gran cluster que podríamos decir que es la clase alta que va a partir del 1’200.000. Esto no representa la realidad del país puesto que corresponde casi al salario mínimo.



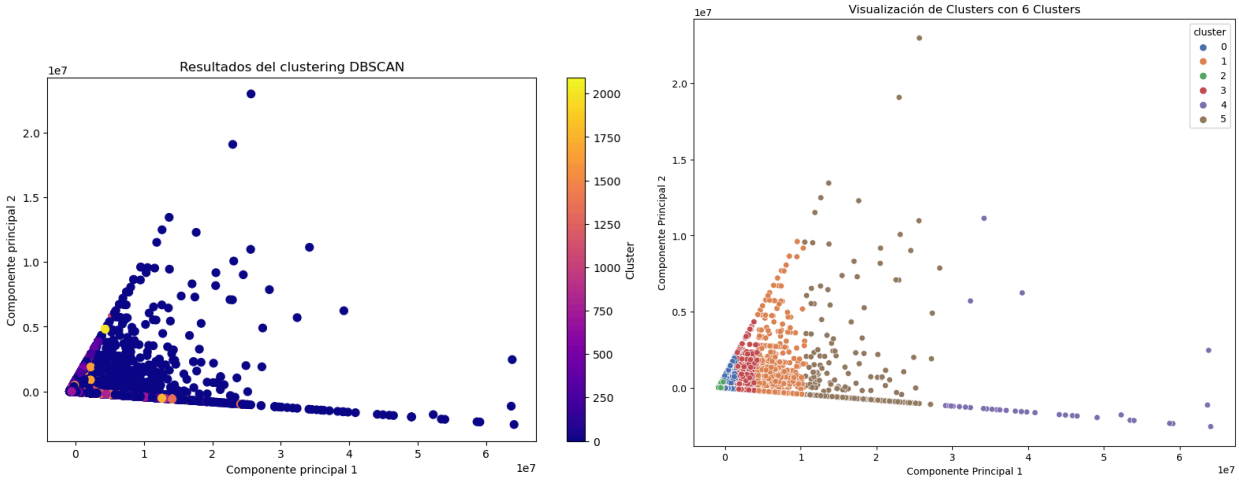
Teniendo una visión inicial de los cluster formados, se decidió utilizar PCA que, como vimos en el punto anterior, únicamente toma en cuenta las variables de ingresos e ingresos por arriendos. De

igual forma se decidió calibrar KNN para encontrar el número de clusters óptimo. Utilizando el método del code podemos esperar que el número óptimo se encontraría entre 5 y 7 clusters.

De esta forma se procedió a graficar cada una de las opciones, notando que a mayor número de clusters se dejaba de dividir la clase media a alta, sino que se encontraban nuevos clusters en la clase baja. Esto debido a que la mayoría de datos se encuentran en este sector.

Teniendo el cuenta el desafío que presentaban los grandes outliers que se tenían en los datos, se decidió utilizar DBSCAN gracias a su capacidad de formar clusters sin tener en cuenta los outliers. Esto podría ser un problema puesto que deseamos ubicar a todas las personas en alguna clase social, pero se decidió hacer el experimento asumiendo que los outliers corresponderían unicamente a la clase alta. Sin embargo, al realizar DBSCAN, luego de haber calibrado los parámetros de eps y min_samples, se obtuvo una granularidad muy alta, dando como resultado más de 2000 clusters. Lo que lo hace infactible para este problema.

De la misma forma se utilizó clustering jerarquico, pero sus resultados respaldaban lo encontrado con KNN. Para este parte se realizó feature selection con Random Forest, el cuál encontró importancia en otras variables fuera del ingreso, aunque el ingreso fuera la principal. Aún así, los clusters formados correspondian a los formados con KNN. De esta forma se decidió continuar con KNN, pero utilizando seis clusters. Esto debido a que es el punto que mejor logra capturar los diferentes grupos que tiene la clase baja sin llegar a sobre dividirlos. Puede que no haya logrado encontrar nuevas divisiones en la clase alta, pero si nos trae una gran pregunta al visualizar que en la clase media propuesta por el DANE caben 3 grupos sociales encontrados, donde se combina clase alta, media y vulnerable. Lo que puede indicarnos que la clasificación del DANE no logra capturar la realidad de los ingresos de la población Colombiana.



Clase	KNN	DANE
1	<85K	$420.676
2	<140K	
3	<400K	
4	**<1M**	$420.676 y $781.000
5	**<2’8M**	**$782.292 - $4’201.570**
6	**>2’9M**	> $4’201.570

En resumen, encontramos una forma de clasificar las clases sociales en Colombia, donde se propone encontrar nuevas divisiones en la clase más bajas (clase vulnerable y clase baja) con el objetivo de entender mejor las diferencias que pueden tener de un nivel de ingresos a otro y proponer políticas públicas con esos resultados.

Conclusiónes

El uso de aprendizaje no supervisado en estudios sociales ha estado incursionando en los últimos años, donde la mayor parte de sus aplicaciones termina utilizando algún algoritmo de *clustering* para su análisis, ya sea estudios de dinámicas urbanas o estudios regionales (Wang, 2022). En este caso, se exploraron diferentes algoritmos como KNN, DBSCAN y clustering jerarquico, llegando en dos de ellos a los mismos resultados. Encontrando así puntos importantes sobre la formación de clases sociales en Colombia.

Primero, se observa que entre más clusters o clases quiera formarse, más se fragmenta la clase baja, pero no separa de la clase media y arriba. Lo que puede mostrarnos que p**ara describir la realidad de Colombia hace falta caracterizar mejor las condiciones de vida de las personas en sus clases más bajas que en las más altas.**

Los entes creadores de políticas públicas deberían diseñarlas de formas más incluyentes. La clase baja contiene a la mayor parte de la población y dentro de ella se tienen varios matizes para caracterizar su nivel de vida. Por lo tanto, se deben generar ayudas y **programas que se enfoquen en aumentar los ingresos de los hogares**, puesto que esto sí hace parte de lo que clasifica finalmente a una persona en un estrato social. Estas políticas pueden ser de educación, emprendimiento o trabajo. Más que subsidios y ayudas, se requiere formar a personas que puedan ganar más dinero y así mejorar la calidad de vida de sus familias.

Cabe resaltar que este fue un proyecto realizado con una muestra de 186,550 datos en un país de más de 18 millones de hogares para el año 2023. Por lo tanto, es insuficiente la muestra para realizar un análisis final.

Bibliografía

DANE (Departamento Administrativo Nacional de Estadística). (2022) *Medición de Pobreza Monetaria y Desigualdad 2022. Colombia: Personas*. https://microdatos.dane.gov.co/index.php/catalog/804/data-dictionary/F16?file_name=Personas

Wang, J., & Biljecki, F. (2022). Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, 129, 103925. https://doi.org/10.1016/j.cities.2022.103925

Ganasegeran, K., Abdul Manaf, M. R., Safian, N., et al. (2024). How socio-economic inequalities cluster people with diabetes in Malaysia: Geographic evaluation of area disparities using a non-parameterized unsupervised learning method. *Journal of Epidemiology and Global Health*, 14(1), 169–183. https://doi.org/10.1007/s44197-023-00185-2