Clustering de Clases Sociales en Colombia

Grupo 8 (MIAD – Aprendizaje No Supervisado)

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

Resumen

En Colombia, la clasificación de la población en clases sociales es un tema de gran debate. El Departamento Administrativo Nacional de Estadística (DANE) ha establecido una metodología para categorizar a las personas basandose en su nivel de ingresos, pero esta ha generado controversia debido a las percepciones sobre los rangos salariales asignados a cada clase. El objetivo de este proyecto es utilizar técnicas de aprendizaje no supervisado, como K-means, clustering jerárquico y DBSCAN, para identificar grupos naturales dentro de la población y compararlos con la clasificación actual del DANE, buscando ofrecer una perspectiva más precisa y representativa de la estructura social colombiana.

A través de un análisis inicial de los datos del DANE, se ha observado que la mayoría de las personas se concentran en ingresos por debajo de los 2 millones de pesos y que las horas trabajadas no se correlacionan directamente con mayores ingresos. El proyecto propone un proceso de preprocesamiento de datos y estandarización de variables, seguido de la implementación de varios algoritmos de clustering. Estos algoritmos permitirán evaluar la estabilidad y robustez de los grupos identificados, y los resultados serán comparados con las clasificaciones oficiales para identificar posibles diferencias y contribuir a una mejor comprensión de las dinámicas socioeconómicas en el país.

Introducción

La clasificación de las clases sociales es un tema de gran relevancia para el diseño y la evaluación de políticas públicas. En Colombia, el DANE ha establecido una metodología para categorizar a la población en clases sociales, basándose principalmente en criterios de ingreso. Sin embargo, esta clasificación ha generado un amplio debate público, especialmente en relación con los rangos salariales considerados para cada clase. Considerando la clase media en un rango de $780.292 a $4′201.570, y la clase alta superior a este.

En este contexto, surge la pregunta: **¿Qué características definen realmente las clases sociales en Colombia?** ¿Coinciden las clasificaciones establecidas por el DANE con los grupos que se formarían al aplicar técnicas de aprendizaje no supervisado?

Se busca utilizar algoritmos de aprendizaje no supervisado para descubrir patrones y agrupamientos naturales en los datos socioeconómicos de la población colombiana. Al comparar

estos resultados con la clasificación de clases sociales del DANE, se espera ofrecer nuevas perspectivas sobre la estructura social del país y evaluar la adecuación de los criterios actuales.

Este proyecto se enmarca dentro del área de **clustering** o agrupamiento. Los algoritmos de clustering buscan identificar grupos de observaciones que sean similares entre sí y diferentes de otras observaciones, sin tener una variable de salida predefinida.

El estudio está motivado por las limitaciones de la clasificación del DANE, que ha sido cuestionada por no reflejar con precisión la realidad social. Los algoritmos de clustering podrían identificar agrupamientos más representativos, lo que podría tener implicaciones importantes para las políticas públicas, permitiendo diseñar programas más específicos y eficaces. Los principales beneficiarios potenciales incluyen el DANE, gobiernos locales y ONG, interesados en ajustar su enfoque hacia los grupos más vulnerables.

Estado del Arte

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 Malasia, se ha estudiado el riesgo de padecimiento de diabetes de variables socioeconómicas (Ganasegeran, 2024). Este estudio utilizó técnicas de agrupamiento para identificar grupos de áreas con patrones similares de carga de diabetes y características socioeconómicas. En particular se utilizaron algoritmos de *agrupamiento jerárquico*, que se resaltan como valiosos a la hora de identificar grupos de áreas con características socioeconómicas y demográficas similares. Estos grupos pueden ser útiles para los profesionales de la salud pública para diseñar intervenciones dirigidas a las áreas con mayor riesgo de diabetes.

Del mismo modo, en Brasil se uso del aprendizaje no supervisado para estratificar el riesgo de parto prematuro utilizando datos socioeconómicos (Lopes, 2022). En este estudio se discuten los métodos de aprendizaje como *k-means*, análisis de componentes principales (*PCA*) y agrupamiento espacial basado en densidad de aplicaciones con ruido (*DBSCAN*). Se identificaron cuatro grupos con altos niveles de incidencia de parto prematuro y tres con niveles bajos. Los grupos con altos niveles de riesgo se componían principalmente de municipios con niveles más bajos de educación, peor calidad de los servicios públicos, como el saneamiento básico y la recolección de basura, y una afro. Los resultados indican una influencia positiva de la calidad de vida y la oferta de servicios públicos en la reducción del riesgo de parto prematuro.

Identificar clases sociales es entonces el punto inicial para diseñar políticas públicas que impacta directamente la calidad de vida de las personas, y así mismo aspectos relacionados a la salud. Sin embargo, el método utilizado en Colombia parece en un inicio no contemplar variables socioeconómicas importantes, dando así una clasificación poco congruente con el nivel de vida.

De esta manera, se propone utilizar algunas técnicas de aprendizaje no supervisado para generar estos grupos, caracterizarlos y compararlos con la clasificación hecha por el DANE.

Descripción de Datos

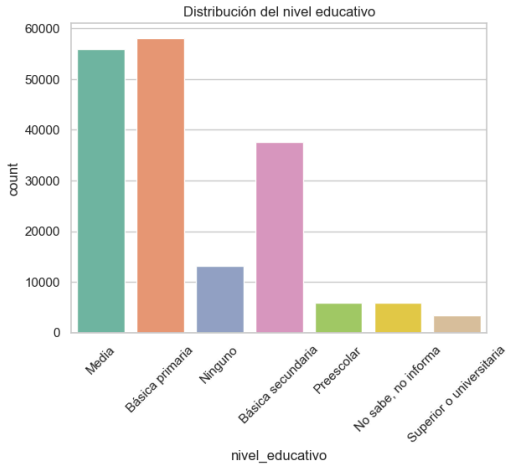
Para el desarrollo del presente proyecto se utilizaron datos abiertos del DANE. En particular se utilizó el dataset de **Pobreza Monetaria y Desigualdad 2022**. Que contiene 132 variables relacionadas con el nivel de ingresos, información sociodemográfica, económica y ocupación para las personas. Se extrajo una muestra de este dataset, contando con *240.001* registros.

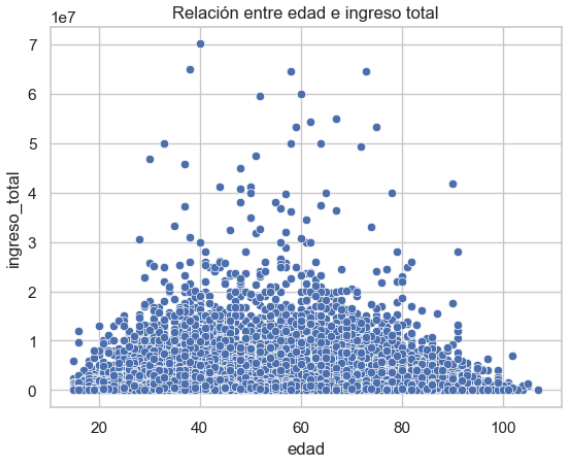
De forma inicial se mapearon los nombres de las columnas codificadas siguiendo el diccionario de datos provisto por el DANE (DANE, 2022). Y de esta forma se generaron estadísticas descriptivas e información sobre los datos faltantes:

Descripción	Observaciones
Datos faltantes	La mayoría de columnas tiene datos faltantes superiores a los 100.000 datos La variable *ingreso_total* es crucial para el análisis y presenta 53,451 valores faltantes. Contando solamente con 186.550 datos Las columnas más completas son *ocupacion*, *nivel_educativo*, *sexo* y *edad*
Sobre el ingreso	El valor promedio de ingreso total es: **$928032.11**, que se encuentra por debajo del salario mínimo actual. La mayoría de personas (quartil 75%) tienen ingresos inferiores a **$1’200.000** Se encuentran salarios por encima de los 7 millones de pesos, pero con muy pocas observaciones. Esto da cuenta de la diferencia de ingresos entre los outliers que se encuentran.
Distribución de variables numéricas	La mayoría de la población se concentra en el rango de 15 a 60 años, con un pico en la adolescencia y otro en la mediana edad. Las horas trabajadas a la semana pueden superar las 120 horas, lo que sería más de 17h por día en promedio. No se observa una relación entre mayor cantidad de horas y mayores ingresos.
Distribucion Variables categoricas	Parece que la distribución entre hombres y mujeres es equilibrada La mayoría de personas viven en la zona urbana Las personas normalmente no reciben beneficios extras en su trabajo además del salario. Existe una gran cantidad de personas cuyo nivel educativo fue la educación media o básica primaria.

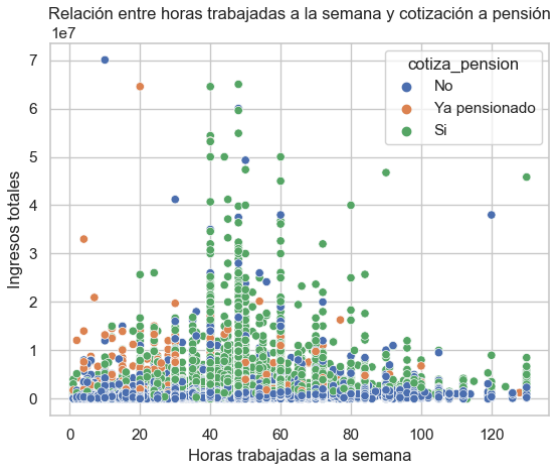
Partiendo de este análisis inicial se realizaron algunos gráficos para entender el comportamiento de las variables a utilizar:

En primer lugar se observa una concentración de los ingresos por debajo de los 2 millones de pesos que parece distribuirse uniformemente respecto a la edad. Sin embargo se observan outliers que pueden superar los 7 millones de pesos. En contraste, la mayor parte del nivel educativo de las personas no llega a la educación superior universitaria. Aunque esto parece no influir en el nivel de ingresos.





De manera similar se decide observar la incidencia de las horas trabajadas a la semana, los ingresos totales y clasificación entre si cotiza pensión o si ya es pensionado. Esto con el objetivo de retratar la informalidad de los oficios en Colombia y su relación con en nivel de ingresos. Se puede observar que el nivel de ingresos más alto se presenta para horas a la semana por debajo de las 20 horas. Sin embargo, la mayor cantidad de personas tiene un ingreso por debajo de los 2 millones de pesos, concentrando la mayor cantidad de personas con trabajo informal en los niveles de ingresos más bajos. Con horas a la semana trabajadas superiores a las 60 horas. Cabe resaltar que se observa gran cantidad de pensionados aún se encuentra trabajando.



Propuesta Metodológica

Para abordar el problema de clasificación de clases sociales, se propone utilizar inicialmente el algoritmo de *K-means* debido a su simplicidad y eficiencia en la creación de clusters cuando se conoce de antemano el número de grupos a formar. K-means es particularmente adecuado para este problema porque nos permite explorar la división de la población en las **cuatro clases sociales** que ha definido el DANE. Sin embargo, somos conscientes de que *K-means* tiene limitaciones, especialmente en la detección de clusters no esféricos y en la sensibilidad a la inicialización de los centroides. Se espera además encontrar dificultades por los outliers que se han identificado.

Para complementar el análisis y asegurar una robusta identificación de patrones en los datos, se considerarán también otros métodos de agrupamiento:

Clustering Jerárquico: Este método se utilizará como una alternativa para explorar la estructura interna de los grupos formados. Esto es útil para identificar las divisiones más sutiles que pueden existir dentro de cada clase social.

Clustering Basado en Densidad (DBSCAN): Será implementado para identificar patrones no lineales en los datos, que podrían no ser capturados por *K-means* o *clustering jerárquico*. Este algoritmo es especialmente efectivo para encontrar grupos de alta densidad que están separados por regiones de baja densidad. Se debe tener en cuenta que este método no clasifica a todas las personas, porque intenta ser resiliente al ruido o los outliers, y para el presente caso de estudio se requiere que cada uno tenga una clasificación. Sin embargo, se decide tenerlo como alternativa para explorar sus resultados.

Se propone entonces el siguiente procedimiento para pasar a las métricas:

	Tareas a realizar
Preprocesamiento de datos	Limpieza de datos y manejo de valores faltantes. Estandarización de las variables para asegurar que todas tengan la misma influencia en el agrupamiento.
Aplicación de algoritmos	Implementación del algoritmo de K-means para la creación inicial de clusters. Evaluación de la estabilidad de los clusters mediante la variación de la inicialización de los centroides. Aplicación del clustering jerárquico para explorar subgrupos y validar los resultados obtenidos con K-means. Implementación de DBSCAN para la detección de patrones no lineales en los datos.

Referencias Bibliográficas

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

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

Lopes, L. M., Barbosa, R. d. M., & Fernandes, M. A. C. (2022). Unsupervised learning applied to the stratification of preterm birth risk in Brazil with socioeconomic data. *International Journal of Environmental Research and Public Health*, 19(9), 5596. https://doi.org/10.3390/ijerph19095596

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