**CLUSTERING SOCIODEMOGRÁFICO PARA ESTUDIANTES ICFES**

*Avance Semana 6 - Aprendizaje No Supervisado – Grupo 6*

**Integrantes:**

David Romero Acosta (da.romeroa@uniandes.edu.co)

Dayana Ortega Leguía (d.ortegal@uniandes.edu.co)

Edgar Garcia Morantes (es.garciam@uniandes.edu.co)

Francisco Martino Gonzalez (f.martino@uniandes.edu.co)

1. **RESUMEN**

Este Proyecto busca segmentar a la población de estudiantes de bachillerato en cuanto a sus posibles preferencias de carrera en educación superior, a través de técnicas de clusterización tomando como base los puntajes para cada área de conocimiento en las pruebas ICFES Saber 11 y la información sociodemográfica capturada. Lo anterior permitiría predecir por cuál carrera universitaria se inclinaría cada clúster de estudiantes.

El examen Saber 11 evalúa a la población estudiantil en términos de 5 áreas de conocimiento, lectura crítica, matemáticas, ciencias naturales, ciencias sociales y ciudadanas e inglés. La base de datos utilizada en este estudio compila la información de 15.528 estudiantes colombianos que presentaron el examen en el primer semestre de 2021 para los cuáles se tienen los puntajes obtenidos en cada una de las áreas de conocimiento y su información sociodemográfica comprendida por género, edad, etnia, ubicación, estrato socioeconómico, acceso a internet, dedicación al estudio, entre otros aspectos relevantes.

1. **RESULTADOS**

La encuesta sociodemográfica comprende 78 variables, de las cuáles se escogieron 14 incluyendo los puntajes en cada área de conocimiento y el puntaje global, así como a criterio de los autores las variables sociodemográficas que soportarían también la clusterización de la población.

Dentro de la base de datos las únicas variables numéricas identificadas corresponden a los puntajes para las 5 áreas de conocimiento y el puntaje global, el resto de las variables son de tipo categóricas por lo tanto las variables seleccionadas se transformaron con encoders y para el manejo de datos faltantes se utilizó la imputación simple.

A continuación, se identifican las 14 variables categóricas identificadas como relevantes para estudio:

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ESTU\_GENERO 12965 non-null object

1 ESTU\_FECHANACIMIENTO 12965 non-null object

2 PERIODO 12965 non-null int64

3 ESTU\_TIENEETNIA 12965 non-null object

4 ESTU\_DEPTO\_RESIDE 12965 non-null object

5 FAMI\_ESTRATOVIVIENDA 12965 non-null object

6 FAMI\_TIENEINTERNET 12965 non-null object

7 ESTU\_HORASSEMANATRABAJA 12965 non-null object

8 PUNT\_LECTURA\_CRITICA 12965 non-null int64

9 PUNT\_MATEMATICAS 12965 non-null int64

10 PUNT\_C\_NATURALES 12965 non-null int64

11 PUNT\_SOCIALES\_CIUDADANAS 12965 non-null int64

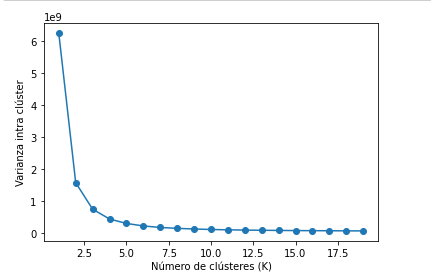
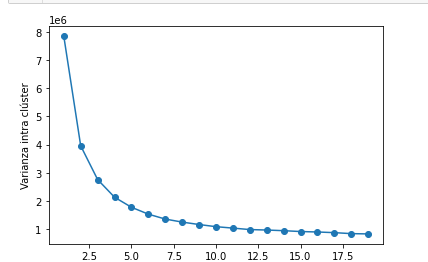
12 PUNT\_INGLES 12965 non-null float64

13 PUNT\_GLOBAL 12965 non-null int64

Se utilizó 2 tipos de técnicas/algoritmos dentro del ámbito del Aprendizaje de Máquinas no Supervisado:

* **Algoritmo de reducción de dimensiones (PCA):** es importante considerar la reducción de dimensiones para poder aplicar mejor el clustering. Utilizamos la técnica de PCA con 2 componentes y obtenemos una varianza explicada cercana a 0.99, lo cual nos permitirá no perder información relevante al momento de clusterizar.
* **Algoritmo de clusterización:** utilizamos inicialmente Kmeans y Kmedoides y vemos, usando la gráfica de varianza intra cluster, que podríamos tomar un valor de k=4 (para ambos algoritmos)

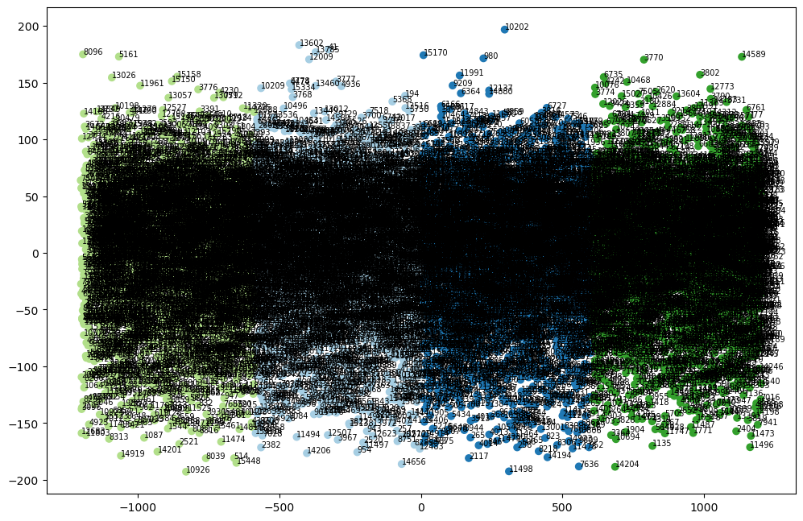
A continuación se observan los gráficos de varianza intra-cluster para ambos algoritmos:



Y vemos que según los indicadores, el Kmedoides tiene un mejor desempeño:



Finalmente vemos los resultados preliminares de la clusterización (donde se identifican los 4 clusters según los dos componentes principales en los que redujimos el dataset):

****

1. **CONCLUSIONES Y RECOMENDACIONES**

**BIBLIOGRAFÍA**

Data Icfes. (s/f). Recuperado el 22 de agosto de 2022, de Icfes website: <https://www2.icfes.gov.co/web/guest/data-icfes>

Ruiz Escorcia, R. R., Arévalo Medrano, J. B., Morillo, G. P., & Acosta-Humánez, P. B. (2017, noviembre 10). Análisis de componentes principales aplicado a la prueba estatal Colombiana Saber 11 Principal component analysis applied to the state Colombian test ICFES Saber 11. Recuperado el 4 de septiembre de 2022, de Revistaespacios.com website: <https://www.revistaespacios.com/a18v39n10/a18v39n10p01.pdf>

Fontecha Ariza, C. (s/f). Análisis de los resultados de las pruebas SABER e I.C.F.E.S. en los componentes de matemáticas y lenguaje y su efecto en los estándares de calidad de la educación en los colegios oficiales de las localidades de Usaquén y Ciudad Bolívar de Bogotá en los periodos 2005 y 2007 (UNIVERSIDAD DE SAN BUENAVENTURA). Recuperado de <http://biblioteca.usbbog.edu.co:8080/Biblioteca/BDigital/65907.pdf>

Pérez-Pulido, M. O., Aguilar-Galvis, F., Orlandoni-Merli, G., & Ramoni-Perazzi, J. (2016). Análisis estadístico de los resultados de las pruebas de estado para el ingreso a la educación superior en la Universidad de Santander, Colombia - Statistical analysis of the results of state tests for admission to higher education at the University of Santander, Colombia. Revista científica, 4(27), 328. doi:10.14483/udistrital.jour.rc.2016.27.a3

Pérez Rubio, Bolívar Atuesta, S., & Correal Núñez, M. E. (2012). Factores que explican el rendimiento escolar de los estudiantes en Bogotá en la Prueba Saber 11 de 2009. Uniandes.

Restrepo Suescún, Correal Núñez, M. E., & Iannini Botero, E. (2012). Análisis de factores asociados al resultado de la prueba Saber 11° del segundo semestre del 2011. Uniandes.

Wikipedia contributors. (s/f). ICFES. Recuperado el 22 de agosto de 2022, de Wikipedia, The Free Encyclopedia website: https://es.wikipedia.org/w/index.php?title=ICFES&oldid=145224997