**Avance Semana 5 - Aprendizaje No Supervisado – Grupo 6**

**Integrantes:**

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

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

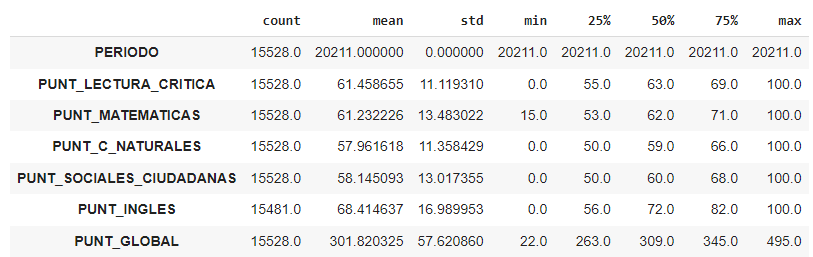
Edgar Garcia Morantes (es.garciam@uniandes.edu.co)

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

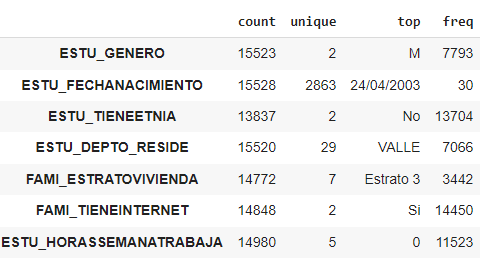
1. **Título del Proyecto:** Clustering Sociodemográfico para Estudiantes ICFES
2. **Estadísticas Descriptivas de los Datos “Limpios”**

Se tomó la base del icfes de los estudiantes que presentaron el examen en el primer semestre de 2021 [2]. Base que cuenta con 15528 filas y 78 columnas.

Para datos numéricos



Para datos categóricos

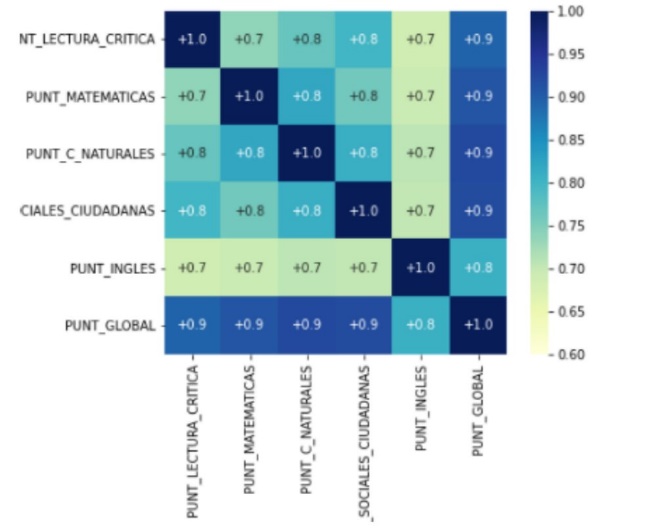


Por otro lado, al revisar algunas relaciones entre variables se observa una correlación entre el estrato y el puntaje obtenido en la prueba. Es decir, a menor estrato, menor calificación media e inversamente, a mayor estrato, mejores calificaciones promedio.

Chart, histogram

Description automatically generated

En cuanto al examen, se observa una alta correlación entre las notas altas de una materia en particular con una nota alta en el puntaje global. También se observa una baja correlación entre las notas de español y matemáticas, y entre las notas de inglés con el resto de las materias.



Finalmente, para la preparación de los datos se realizan las siguientes actividades con el fin de poder utilizar toda la información disponible posible:

* Transformación de variables categóricas con Encoders.​
* Reducción de dataset original filtrando por las variables relevantes escogidas en la exploración.
* Como estrategia de manejo a datos faltantes utilizamos la imputación simple.

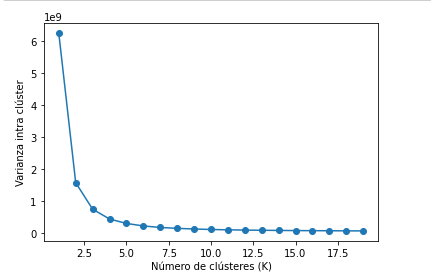
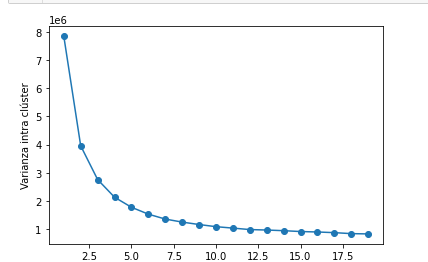
1. **Descripción del Algoritmos a utilizar**

Utilizamos 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)

1. **Resultados Preliminares**

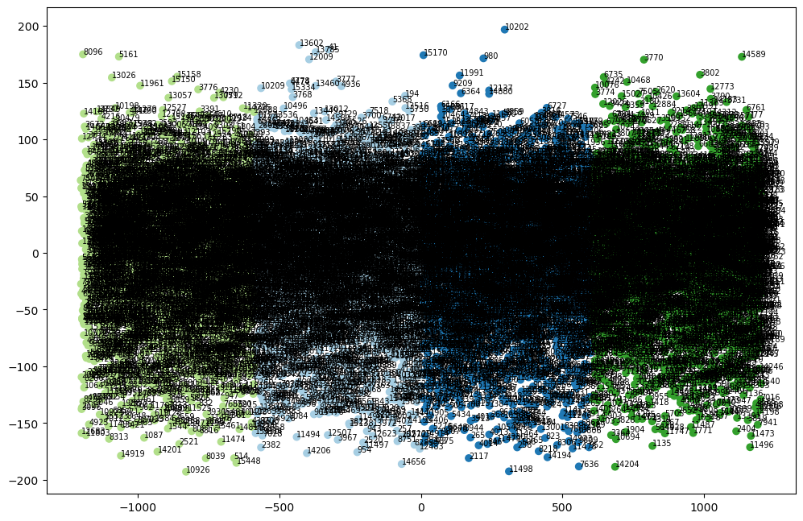
Podemos ver los gráficos de varianza intra-cluster para ambos algoritmos:



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. **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