# CLUSTERING SOCIODEMOGRÁFICO PARA ESTUDIANTES ICFES

Dayana Ortega, David Romero, Edgar Garcia, Francisco Martino

25 de septiembre de 2022

## 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.

## 2. Introducción

### 2.1. Objeto de investigación

El presente analisis busca responder la pregunta: ¿Cuáles son los factores más relevantes que permiten segmentar el grupo de estudiantes que presenta la prueba ICFES Saber 11 con el fin de saber a qué carrera de educación superior se puede presentar? y tiene como cliente potencial a las Universidades y estudiantes que quieren escoger una carrera de educación superior.

### 2.2. Contexto

El ICFES [2] (Instituto Colombiano para la Evaluación de la Educación) es una entidad que evalúa estudiantes para el acceso a la educación superior y que realiza investigaciones de la calidad educativa con el fin de encontrar factores que inciden en esta. Por lo tanto, en este estudio se quiere segmentar a los estudiantes con el fin de hacerle una recomendación de carrera en una institución de educación superior a partir de sus resultados.

Debido a que no se cuenta con datos con una respuesta supervisada, es decir, solo se tiene la información de los predictores, este problema hace parte de los modelos de aprendizaje no supervisado. El problema pertenece a una combinación de tareas de reducción de dimensión y clustering. Esto debido a que se dispone de una base de datos de 82 variables y será necesario poder reducirlas para evitar impactos negativos en los resultados; por otro lado, la tarea central del problema es clustering, dado que se prioriza encontrar los factores más relevantes para segmentar a los estudiantes.

### 2.3. Contraste de literatura

En Colombia algunos analisis[6] similares se han realizado haciendo uso de análisis de componentes principales para inspeccionar los datos de las pruebas de estado del año 2012. La metodología se enfocó principalmente en la reducción dimensional sin realizarsegmentación como se plantea en esta propuesta. Otros[1] se han centrado en poblaciones objetivo tales como la ciudad de Bogotá D.C. entre el 2005 y 2007. Se trata de un estudio principalmente conductual y exploratorio con énfasis en los resultados obtenidos en los componentes de matemáticas y lenguaje. Se diferencia principalmente en la ausencia de modelado estadístico.

[4]En este estudio se buscaba determinar el riesgo asociado a deserción estudiantil de la universidad de Santander considerando como predictores los resultados de la prueba de estado. Se destaca el uso de algoritmos de clasificación para tratar de entender la relación de dependencia entre variables, sin embargo, difiere del planteamiento propuesto en que se centra en análisis supervisado. También los autores de [3] se han enfocado en en técnicas de regresión de analisis supervisado para predecir el comportamiento de puntaje que tuvieron los estudiantes que participaron en la prueba saber 11. Este enfoque se encuentra bajo la esfera de análisis supervisado.

[5]De forma similar al elemento anterior se utilizaron técnicas de aprendizaje estadístico supervisado para modelar los factores que inciden en el rendimiento escolar en la ciudad de Bogotá usando como banco de datos los resultados obtenidos en la prueba saber 11 del año 2009.

# 3. Materiales y Métodos

## 3.1. Fuente de datos y preparación

Se tomó la base del icfes de los estudiantes que presentaron el examen en el primer semestre de 2021 [2]. Esta cuenta con 15528 filas y 78 columnas. Tomando como insumo los resultados de los trabajos realizados anteriormente se realiza un filtrado de las variables disponibles y se clasifican en: datos numéricos y categóricos. Algunas descriptivas pueden encontrarse en la Tabla1 y Tabla2 respectivamente.

## 3.2. Análisis descriptivo

Los consolidados de información para las variables numéricas incluyen las edades de los estudiantes que presentan la prueba Saber 11, el periodo lectivo y los puntajes de las asignaturas Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales Ciudadanas e Inglés; existe también una variable llamada Puntaje Global que es resultante de la combinación lineal de los puntajes individuales de cada materia.

| Variable | Cant. | Media | Desv. | min. | Q1 | Q2 | Q3 | Max. |
|---|---|---|---|---|---|---|---|---|
| Edad | 13971 | 17.15 | 2.93 | 11 | 16 | 17 | 17 | 70 |
| Periodo | 13971 | 20211 | 0.00 | 20211 | 20211 | 20211 | 20211 | 20211 |
| Lectura Crítica | 13971 | 61 | 11.22 | 0 | 54 | 63 | 69 | 100 |
| Matemáticas | 13971 | 60.66 | 13.52 | 15 | 52 | 62 | 70 | 100 |
| C. Naturales | 13971 | 57.49 | 11.39 | 0 | 50 | 58 | 66 | 100 |
| Sociales Ciudadanas | 13971 | 57.63 | 13.05 | 0 | 49 | 59 | 67 | 100 |
| Inglés | 13925 | 67.60 | 17.13 | 0 | 55 | 70 | 81 | 100 |
| Puntaje Total | 13971 | 299.19 | 57.99 | 22 | 260 | 360 | 343 | 495 |

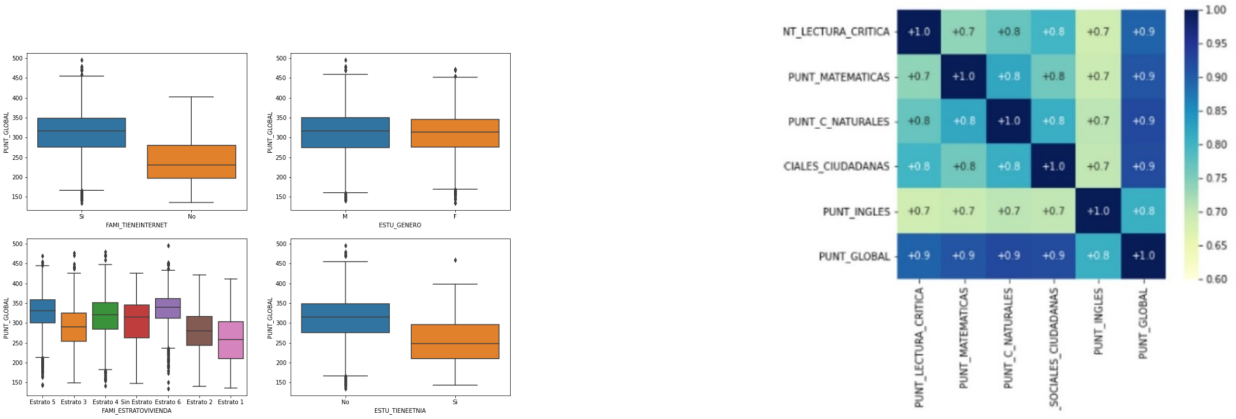
Tabla 1: Descriptivas datos numéricos

| Variable | Cantidad | Cardinalidad | Moda | Frecuencia |
|---|---|---|---|---|
| Género | 13966 | 2 | M | 7026 |
| Tiene Etnia | 12289 | 2 | No | 12157 |
| Departamento | 13963 | 29 | Valle | 6263 |
| Estrato Vivienda | 13267 | 7 | Estrato 3 | 3173 |
| Tiene Internet | 13342 | 2 | Si | 12952 |
| Horas Trabajo Semanal | 13471 | 5 | 0 | 10261 |

Tabla 2: Descriptivas datos categóricos

Por su parte las variables categóricas preseleccionadas incluyen el género binario del estudiante, indicativo de pertenencia a una etnia, departamento geográfico de residencia, estato socioeconómico e indicativos de acceso a internet y simultaneidad con empleo.

Al revisar algunas relaciones en la Figura1a 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.




(a) Diagramas de Caja

(b) Correlaciones

Figura 1: Puntaje Global vs categóricas y correlaciones entre puntajes

En cuanto al examen, se observa una alta correlación en la Figura1b 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.

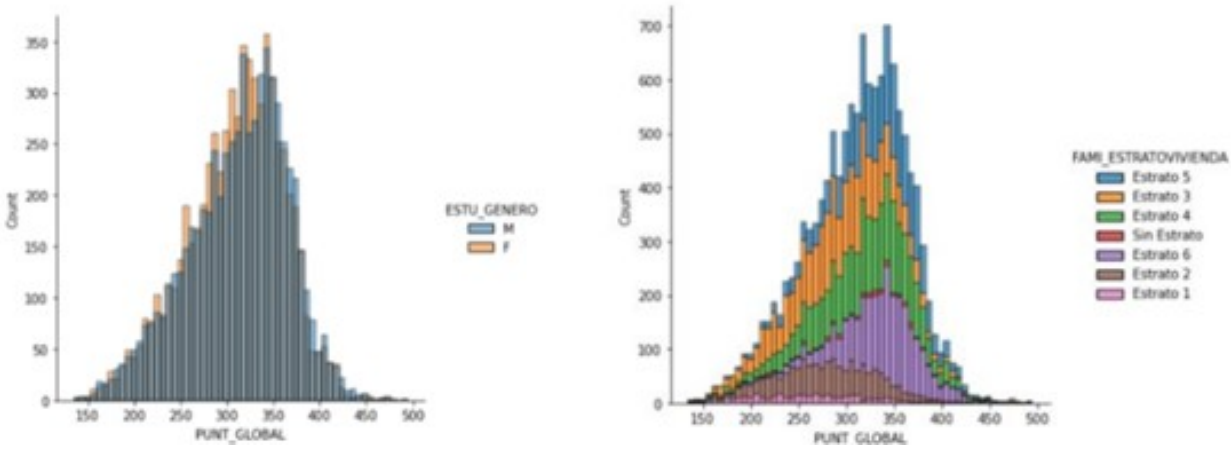



Figura 2: Distribuciones de puntajes según variables

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.

### 3.3. Descripción del algoritmo

Las dos técnicas/algoritmos utilizadas pertenecen al 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. La primera component principal explica un 71.1 % de la varianza. Esta le da mayor peso a las variables Departamento, estrato y a los puntajes de los resultados de las pruebas de Lectura, Matemáticas, Naturales, Sociales e Inglés. De manera similar, la segunda componente principal le da mayor importancia a las variables indicadas anteriormente adicionando la variable edad; se obtiene así una reducción dimensional en la que se explica aproximadamente el 84 % de la varianza del conjunto de datos.
- Algoritmos de segmentacion (clustering): se prueban los algoritmos Kmeans, Kmedoides y el modelo basado en densidad DBSCAN.

# 4. Resultados y Discusión

## 4.1. Implementación de algoritmos

En los primeros dos algoritmos (KMeans y K Medoids) se encuentra similitud en sus métricas de varianza intra cluster e índice de Silhoutte de la Figura 3 en los que se estima un valor de k=2 (para ambos algoritmos) es adecuado.

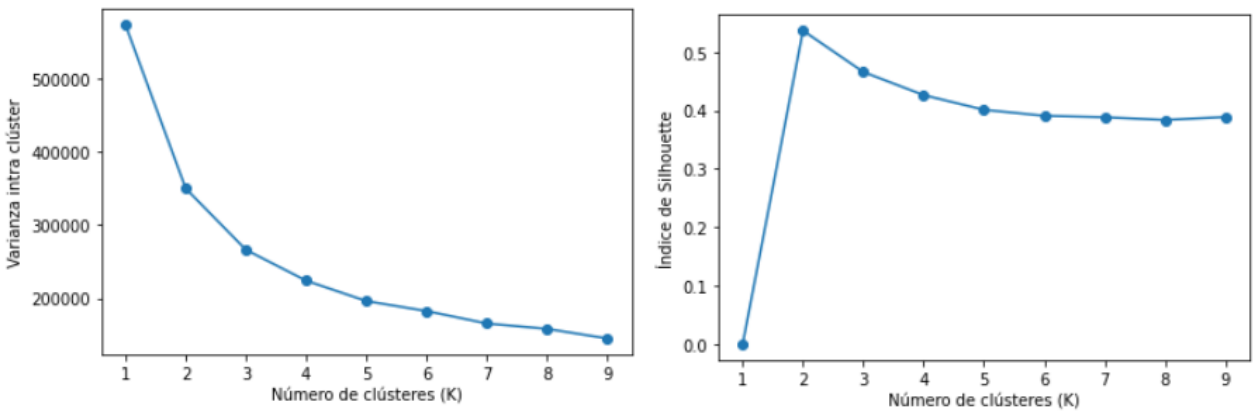


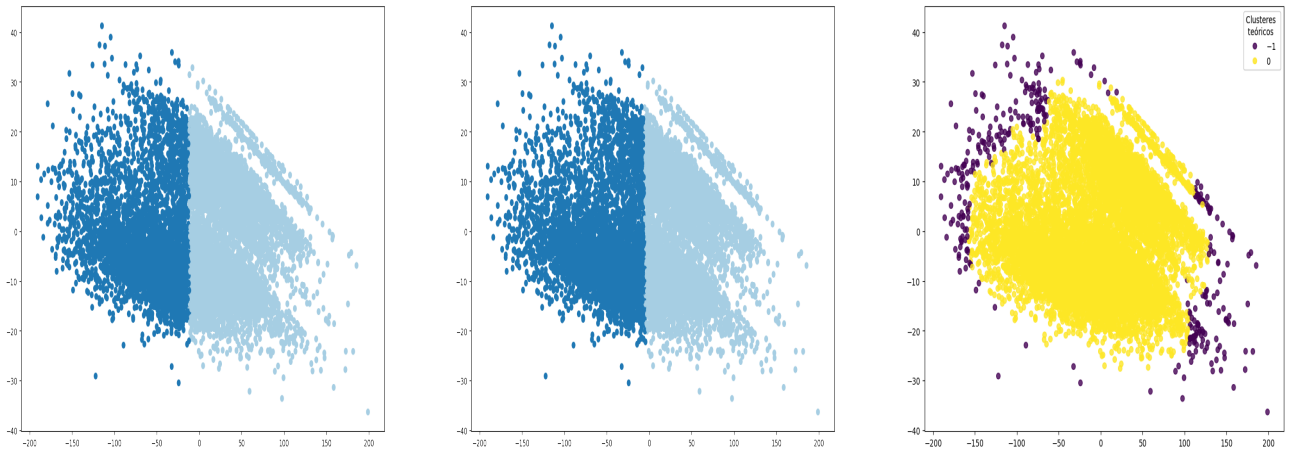

Figura 3: Varianza Intracluster e Índice de Silhouette

Por su parte el algoritmo basado en densidades establece como resultados adecuados un eps=6.0 y una cantidad minima de muestras de 30 elementos.

## 4.2. Resultados obtenidos

Los resultados que se muestran en la Figura 4 muestran que no hay diferencia significativa en las segmentaciones realizadas por los algoritmos de KMeans y KMedoids, principalmente porque no hay figuras fuertemente definidas por los dos componentes principales, en el algoritmo basado en densidades, la clasificación de puntos 'atípicos' se deriva de aquellos que se encuentran en las areas exteriores de la masa central.

Analizando en profundidad estos resultados se encontró que los estudiantes en su mayoría corresponden al departamento del Valle en Colombia, con estrato y condiciones de internet y etnia similares, de acuerdo con lo anterior podemos concluir que la muestra de datos iniciales se encontraba sesgada a un grupo con características muy similares. Se considera que para futuros análisis se debe incorporar mayor cantidad de datos donde se puedan tener estudiantes de todas las regiones del país así como de diferentes cohortes, dado que en este caso se analiza sólo una cohorte de estudiantes que presentó el examen. Con esta información se podría identificar que lo que realmente segmenta con profundidad a los estudiantes podría relacionarse con el departamento o región de donde provienen y la cohorte ala que pertenecen.




(a) KMeans (b) Kmedoides (c) DBSCAN

Figura 4: Visualización de los clústers en cada Algoritmo empleado

## 5. Conclusiones

Debido a la gran cantidad de variables sociodemográficas y de puntaje presente en los datos, la selección de variables y la selección de componentes principales PCA permitieron abordar el problema de aprendizaje no supervisado reduciendo la dimensión de los datos pero aún explicando en más de un 70

Luego de aplicar los algoritmos de K-medias, K-medoides y DBSCAN se logra identificar que a partir de los datos no se logra una segmentación en clústers de los estudiantes, es decir, las características sociodemográficas seleccionadas y los puntajes de sus exámenes no denotan que existan diferentes grupos de estudiantes con características similares. Ahora bien, para esta población analizada no se encuentran diferencias significativas entre los diferentes puntajes para las distintas áreas de conocimiento, entonces no es posible afirmar que existe un clúster de estudiantes con inclinaciones similares hacia un área de estudio específico.

En líneas generales, se considera que el proyecto y sus estudios posteriores pueden tener un impacto positivo en el ecosistema educativo general de Colombia, para poder contribuir en mejorar la igualdad de oportunidades y acceso a la educación de los estudiantes que los ayude a alcanzar mayor éxito y calidad de vida en el futuro.

## Referencias

[1] C. A. Fontecha Ariza. 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. Master's thesis, Universidad de San Buenaventura - Sede Bogotá.

[2] ICFES-Colombia. Data icfes. Available at `https://www2.icfes.gov.co/web/guest/data-icfes` (2022/08/22).

[3] J. K. Perez Rubio. Factores que explican el rendimiento escolar de los estudiantes en bogota en la prueba saber 11 de 2009. Master's thesis, Universidad de los Andes - Bogotá, 2012.

[4] M. O. Pérez Pulido, F. Aguilar Galvis, G. Orlandoni Merli, and J. Ramoni Perazzi. Análisis estadístico de los resultados de las pruebas de estado para el ingreso a la educación superior en la universidad de santander, colombia. *Revista Científica*, 27(1):328–339, 2016.

[5] M. F. Restrepo Suescun. Analisis de factores asociados al resultado de la prueba saber 11° del segundo semestre del 2011. Master's thesis, Universidad de los Andes - Bogotá, 2012.

[6] R. R. Ruiz Escorcia, J. B. Arévalo Medrano, G. P. Morillo, and P. B. Acosta Humánez. Análisis de componentes principales aplicado a la prueba estatal colombiana saber 11. *Revista Espacios*, 39(1):12, 2018.