Propuesta Inicial 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 y Recomendador de Carrera Universitaria para Estudiantes ICFES

2. Resumen

Cada año en Colombia todos los estudiantes de grado 11 en el nivel de bachillerato asisten a una extensa jornada para presentar un examen integral que determina su futuro y puerta de entrada hacia la Educación Superior, sin embargo, se ha evidenciado que condiciones sociales, económicas y demográficas inciden en los resultados de este.

El examen Saber 11 es una evaluación estándar que se realiza semestralmente por el Icfes y tiene como objetivos: ser criterio de selección para la entrada de estudiantes a las Instituciones de Educación Superior, al igual que realizar un monitoreo de la calidad de la formación que ofrecen los establecimientos de educación media y dar información que permite estimar el valor agregado de la educación superior.

A partir de los resultados y la información sociodemográfica capturada por esta institución, se plantea desarrollar el presente proyecto que permita en primer lugar segmentar a los estudiantes utilizando todas las características sociodemográficas y de desempeño disponibles y en segundo lugar generar un modelo de recomendación que sugiera la potencial carrera universitaria en la que el estudiante podría tener éxito. Todo lo anterior mencionado, bajo el enfoque del Aprendizaje de Máquinas no Supervisado. Este informe analiza los datos disponibles para el período 2021 de los resultados del Examen Saber 11.

3. Introducción

Items	Descripción
Pregunta de investigación	¿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?
Motivación	El ICFES (Data Icfes s.f) (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.
Asociación con el ANS*	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 contamos con una base de datos de muchas columnas (en la sección 5 se verá que son 82 dimensiones) y será necesario poder reducirlas para evitar impactos negativos en los resultados; por otro lado, la tarea central del problema es clustering, dado que queremos encontrar los factores más relevantes para segmentar a los estudiantes.
Cliente potencial	Universidades y Estudiantes que quieren escoger una carrera de educación superior

* Aprendizaje No Supervisado

4. Revisión Preliminar de Antecedentes en la Literatura

Referencia	Descripción
(Ruiz Escorcia, R. R, 2017)	El análisis busca responder una pregunta de investigación similar 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 realizar segmentación como se plantea en esta propuesta.
(Fontecha Ariza, C. (s/f).)	El análisis se centra en una población objetivo de 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.
(Pérez-Pulido, M. O., 2016)	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.
(Pérez Rubio, Bolívar Atuesta, et. Al, 2012)	Este trabajo de grado se enfoca 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.

Referencia	Descripción
(Restrepo Suescún, Correal Núñez, 2012)	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.

5. Descripción Detallada de los Datos

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 con los siguientes tipos de datos: int, float, object, string y datetime. Entre las características se encuentra información relacionada con la nacionalidad, el género, el departamento y municipio de residencia, estrato socio económico, datos relacionados al entorno familiar, datos relacionados al colegio del estudiante y el desempeño en cada área de evaluación del examen. Algunas estadísticas descriptivas son:

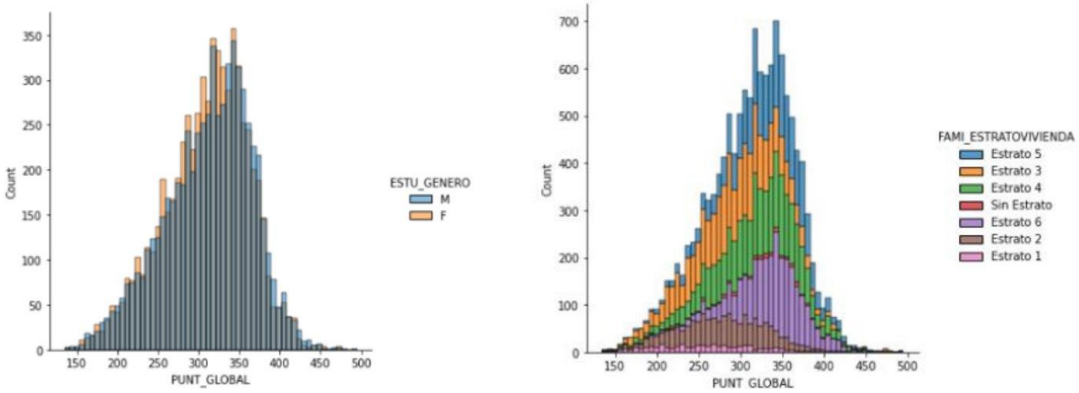
Para datos numéricos

	count	mean	std	min	25%	50%	75%	max
PERIODO	15528.0	20211.000000	0.000000	20211.0	20211.0	20211.0	20211.0	20211.0
PUNT_LECTURA_CRITICA	15528.0	61.458655	11.119310	0.0	55.0	63.0	69.0	100.0
PUNT_MATEMATICAS	15528.0	61.232226	13.483022	15.0	53.0	62.0	71.0	100.0
PUNT_C_NATURALES	15528.0	57.961618	11.358429	0.0	50.0	59.0	66.0	100.0
PUNT_SOCIALES_CIUDADANAS	15528.0	58.145093	13.017355	0.0	50.0	60.0	68.0	100.0
PUNT_INGLES	15481.0	68.414637	16.989953	0.0	56.0	72.0	82.0	100.0
PUNT_GLOBAL	15528.0	301.820325	57.620860	22.0	263.0	309.0	345.0	495.0

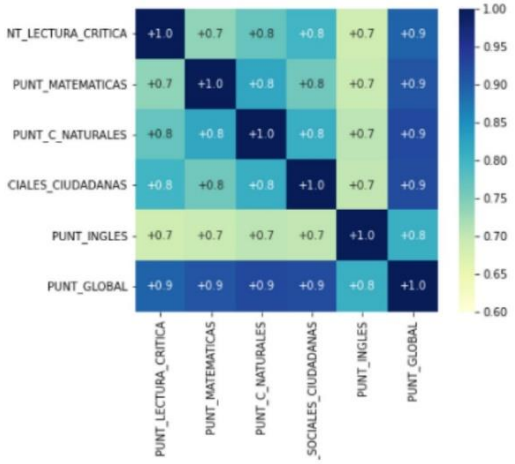
Para datos categóricos

	count	unique	top	freq
ESTU_GENERO	15523	2	M	7793
ESTU_FECHANACIMIENTO	15528	2863	24/04/2003	30
ESTU_TIENEETNIA	13837	2	No	13704
ESTU_DEPTO_RESIDE	15520	29	VALLE	7066
FAMI_ESTRATOVIVIENDA	14772	7	Estrato 3	3442
FAMI_TIENEINTERNET	14848	2	Si	14450
ESTU_HORASSEMANATRABAJA	14980	5	0	11523

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.



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 realizarán las siguientes actividades con el fin de poder utilizar toda la información disponible posible:

- Transformación de variables categóricas a dummies.
- Reducción de dataset original filtrando por las variables relevantes escogidas en la exploración.
- Como estrategia de manejo a datos faltantes se procederá a evaluar varias estrategias con el fin de encontrar el mejor rendimiento en los resultados obtenidos. Exploraremos estrategias tales como la imputación simple y otras más sofisticadas como MCAR y MNAR

6. Propuesta Metodológica

Planteamos utilizar 3 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, dado que como vimos en la sección 5 del presente documento, contamos con 82 columnas y será necesario mitigar el impacto de la dimensionalidad en la implementación de los algoritmos.

- Algoritmo de clusterización (Jerárquico Aglomerativo o DBSCAN): es necesario implementar algoritmos de clusterización para resolver el problema planteado y encontrar los grupos de estudiantes que se parezcan más. Deberemos probar entre diferentes algoritmos, pero la ventaja potencial de un DBSCAN será poder eliminar el ruido; sin embargo, tendremos que evaluar los trade-offs versus la complejidad del dataset, para esto, un algoritmo jerárquico podría ser una buena opción.

- Modelos de recomendación (Filtro colaborativo): para poder resolver el problema de sugerir una potencial carrera universitaria en la cual los estudiantes tengan éxito, utilizaremos técnicas asociadas a los modelos de recomendación. Un candidato a utilizar será el Filtro colaborativo.

7. 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