**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**

Se define el problema/reto al que se enfrenta el equipo, resaltando el resultado principal que se espera del trabajo sin entrar en los detalles. Está escrito para una audiencia general, más no especializada, y describe lo que se planea hacer haciendo explícita la contribución del trabajo. Tiene como máximo 500 palabras. [15 puntos]

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.

1. **Introducción**

Se establece un problema o una pregunta bien definida con una motivación clara para resolverla. Se define quién es el cliente potencial y cuál es el contexto organizacional donde surge el problema o la pregunta de interés, lo que se espera solucionar con el uso del aprendizaje no supervisado, y a qué área del aprendizaje no supervisado pertenece el problema o pregunta a resolver. [15 puntos]

|  |  |
| --- | --- |
| **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 [1] (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

1. **Revisión Preliminar de Antecedentes en la Literatura**

Se hace una revisión de la literatura nacional e internacional donde se exploran preguntas o problemas similares, y se utilizan métodos similares. Se establece las similitudes y diferencias de los enfoques presentados en la literatura con el enfoque propuesto en esta entrega. [10 puntos]

1. **Descripción Detallada de los Datos**

Se menciona la fuente de los datos. Se definen las variables, su tipo, y sus estadísticos descriptivos. Se utilizan tablas y/o herramientas gráficas como diagramas de dispersión, histogramas, correlogramas, etc., para establecer el rol que cumple cada variable en el contexto de la resolución del problema bajo estudio. [25 puntos]

**Exploración de datos preliminar**

A picture containing text

Description automatically generated

**Tipo de Datos​**

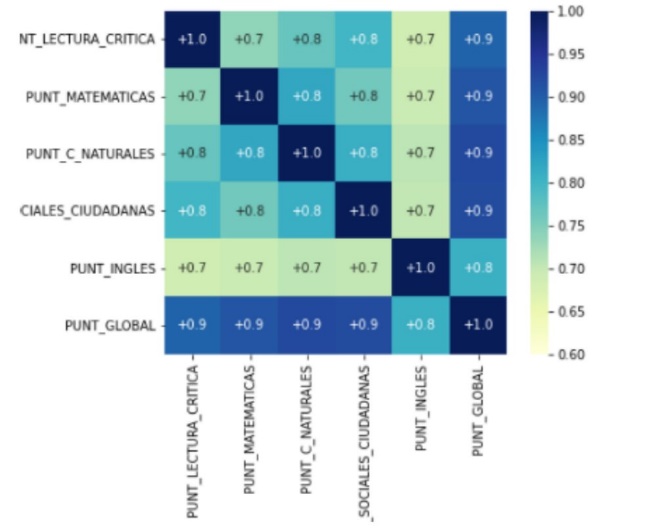
Se cuenta con los siguientes tipos de datos: int, float, object, string y datetime.​

**Descripción base de datos​**

Chart, histogram

Description automatically generatedEsta base consta de 30963 registros y 82 características entre las que se encuentran: 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.​

Se observa una correlación entre el estrato y el puntaje obtenido en la prueba. 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.​

Es prudente eliminar los puntajes por materia para realizar futuras predicciones, puesto que sus combinaciones lineales dan como resultado el puntaje global.​

**Preparación de los datos**

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

* 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 eliminarlos, esto podría resultar en una reducción de 28,95% del conjunto original

1. **Propuesta Metodológica**

Se define un algoritmo/técnica que se planea utilizar y se plantea qué otros algoritmos/técnicas son candidatos a utilizarse para solucionar la pregunta de interés. Se argumenta por qué el elegido es el adecuado dado el contexto del problema y de la organización. [15 puntos]

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.

1. **Bibliografía**

Se citan los artículos mencionados en el texto, usando de forma correcta y consistente el estilo de referencia que se hay escogido usar (Chicago, APA, MLA, etc.). [5 puntos]

1. Wikipedia. (2022, Ago 22). ICFES [online]. Disponible:

<https://es.wikipedia.org/wiki/ICFES>

1. Icfescol. (2022, Ago 22). DataIcfes [online]. Disponible

<https://www2.icfes.gov.co/web/guest/data-icfes>

Se entrega el documento en formato .pdf y no se exceden las 5 páginas, excluidas las referencias. Se utiliza una fuente estándar (Times o Arial) de 11 o 12 puntos. [5 puntos]