**Avance 2 – Propuesta proyecto final del curso de Aprendizaje No Supervisado**

**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. **Elegir una de las preguntas utilizando la retroalimentación recibida en el avance anterior y expandir sobre su motivación a responderla.**

De acuerdo con la retroalimentación recibida se realiza elección del proyecto número 1 presentado.

|  |  |
| --- | --- |
| **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 |
| Cliente potencial | Universidades y Estudiantes que quieren escoger una carrera de educación superior |

\* Aprendizaje No Supervisado

1. **Describir si el problema pertenece a una tarea de reducción de dimensión, clustering, o una combinación de los dos, explicando el por qué.**

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 4 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.

1. **Proponer al menos un algoritmo/técnica de aprendizaje no supervisado que ustedes crean que es la adecuada para responder la pregunta.**

* PCA: para poder reducir dimensiones
* Kmeans o Kmedoides: para clusterizar

1. **Presentar estadísticas descriptivas utilizando tablas y/o visualizaciones de los datos crudos que tengan a su disposición. Describir el plan que van a seguir para tener los datos listos. Por ejemplo, cómo van a limpiarlos, que van a hacer con los datos faltantes, etc.**

**Exploración de datos preliminar**

A picture containing text

Description automatically generated

**Tipo de Datos​**

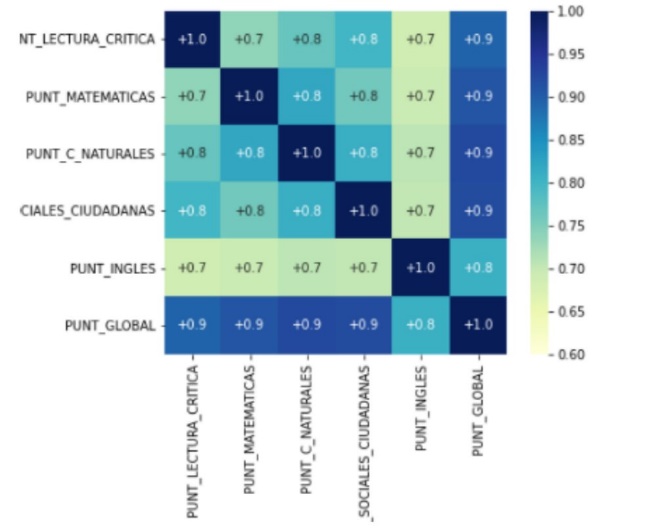
Se cuenta con los siguientes tipos de datos: int, float, object, string y datetime.​

**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. **Con base en los roles definidos en el documento de la semana anterior, delinear las actividades que se llevaron y llevarán a cabo para la primera entrega calificada del proyecto. Ser preciso, y tomar este punto como un contrato entre los miembros del equipo.**

|  |  |
| --- | --- |
| **Integrante** | **Rol** |
| David | Apoyo en la etapa de extracción, transformación y carga de datos de las fuentes de información. |
| Francisco | Apoyo en la formulación de modelos analíticos para resolver la pregunta de investigación. |
| Edgar | Apoyo en las métricas de desempeño y selección del mejor modelo para el problema planteado. |
| Dayana | Identificar oportunidades de aplicación del producto final para generar valor dentro de las organizaciones objetivo. |

En paralelo, todos estaremos apoyando al líder de cada rol y en la edición del informe final del trabajo.

**Referencias**

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>