**Gestión y almacenamiento de datos:**

**Descripción del Dataset escogido**

Alma Brand

Keiry Pereira

Sebastián Hurtado

Simón Ruiz

De acuerdo con la problemática seleccionada, que tiene como objetivo analizar datos que permitan identificar las causas o los elementos que inciden en el proceso de aprendizaje de los jóvenes en Colombia, nuestro grupo ha decidido profundizar en la exploración de estadísticas relacionadas con las pruebas Saber ICFES. Estas pruebas son un indicador crucial para evaluar el nivel educativo de los jóvenes recién graduados de la educación secundaria, además de proporcionar datos socioeconómicos de quienes las presentaron. Esto nos brinda la oportunidad de llevar a cabo un análisis de datos exhaustivo y extraer conclusiones significativas en relación con la influencia de factores sociales y económicos en el aprendizaje de los jóvenes.

El dataset que hemos seleccionado contiene información socioeconómica de todas las personas que presentaron las pruebas ICFES en el año 2019, incluyendo los puntajes obtenidos en cada área evaluada y el puntaje global de cada individuo. Elegimos este dataset por dos razones principales. En primer lugar, el año 2019 es el más reciente en el que el ICFES liberó estos datos para uso público, lo que garantiza su relevancia y actualidad. Además, el año 2019 precedió directamente a la pandemia global, lo que sugiere que los resultados no estarían influenciados por los eventos canónicos que tuvieron lugar en todo el mundo a partir de ese momento, lo que aumenta la precisión de nuestras conclusiones.

Después de adquirir el conjunto de datos, se llevó a cabo un proceso de depuración en el que se eliminaron aquellas variables que no guardaban relevancia con el objetivo de la investigación. Entre estas variables, se incluían datos como el lugar donde la persona realizó la prueba y el tipo de documento de identidad que poseía, entre otros similares.

Posteriormente, se cambiaron algunas variables booleanas que respondían a preguntas de si y no, o de hay o no hay, por un valor numérico binario, esto con el fin de analizar de manera más eficiente el dataset y así procesar de manera más rápida la correlación de los datos con la salida estipulada, que en este caso es el puntaje global.